# Supplementary Material

## I. Methods for computational literature validation of top results

To evaluate the relevancy of top hits from the 'cross-group' statistical distances in comparison to the other three conventional methods, weighted least squares (WLS), single-cell differential expression followed by gene set enrichment (SCDE+Enrichment), and gene set enrichment analysis (GSEA), *PubMed* was searched using the online tool *Pubmatrix* [1] (using method adopted from Yang et al. [2]). The search retrieved the co-occurrence of literature for prostate cancer & treatment versus 187 Pathway Interaction Database (PID) pathways. The 6 prostate cancer & treatment terms specified were *prostate cancer, prostatic adenocarcinoma, prostatic carcinoma, enzalutamide, antiandrogen*, and *androgen deprivation.* For simplification and more complete searches, the PID descriptions were condensed (and sometimes separated) to only informative gene symbols and minimal descriptors (available upon request). There were 198 condensed terms associated with 187 PID pathways. In order to estimate the odds of the co-occurrence of literature for top-hit pathways, it is also required to retrieve the total number of *PubMed* articles for the 198 pathway terms irrespective of the prostate cancer and, similarly, the number of articles associated with the 6 prostate cancer & treatment terms. The keyword "*all [sb]"* was used as the cross-term in *Pubmatrix* to retrieve all counts associated with the key terms (http://www.ncbi.nlm.nih.gov/books/NBK3827/).

Once the counts of co-occurring literature were obtained, 2x2 contingency tables were constructed. Since no method discovered pathways while controlling for multiplicity of testing, a nominal 5% p-value threshold determined the number of top-hit pathways. Each method may differ with respect to the number of top hits in this fashion. To maintain a fair comparison, the minimum number of top hits associated with a method was employed while evaluating all four methods. In our case study, both GSEA and SCDE+Enrichment had the minimal number of hits, 16 pathways. An example 2x2 contingency table is illustrated below:

**Table S1. Exemplar 2x2 table for literature validation of top-prioritized pathways.**

| Cells contain counts of *PubMed* articles retrieved via *Pubmatrix* (*N = 25,843,603* as of 3/17/2016) | Prostate cancer & treatment | NOT Prostate cancer & treatment | Marginal sum |
|---|---|---|---|
| Top-hit pathways | *n1* | *n3 - n1* | *n3* |
| NOT in top-hit pathways | *n2 - n1* | *N - n2 - n3 + n1* | *N – n3* |
| Marginal sum | *n2* | *N - n2* | *N* |

A Fisher's exact test was employed to assess any association of prioritized genes (associated with pathways) with prostate cancer. Odds ratios were estimated in this process, with odds ratios greater than 1 indicating corroboration of the top hits in literature. With the large counts present in this search, small derivations from unity are likely to be statistically significant. To assess whether a given odds ratio is extreme for these data an *ad hoc* resampling procedure was performed. A random set of 16 PID pathways was selected without replacement and the odds ratio was computed. This was repeated 100,000 times to produce an empirical distribution of odds ratios. An empirical one-sided 'p-value' was computed as the proportion of odds ratios greater than the sample odds ratio for a given pathway prioritization method.

**References**

1. Becker KG, Hosack D a, Dennis G, Lempicki R a, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4**:61.

2. Yang X, Huang Y, Crowson M, Li J, Maitland ML, Lussier YA: **Kinase inhibition-related adverse events predicted from in vitro kinome and clinical trial data**. *J Biomed Inform* 2010, **43**:376–384.

## II. Within-patient CTC pairs reveal patient heterogeneity

This section complements the analysis of cross-group and within-group pairings presented in the main article by inspecting within-patient statistical distances and significant pathway dysregulation.

**Table S2. Summary of statistical distances within the five resistance-associated pathways identified through the 'cross-group' analysis.** Within-patient pairs of CTCs were compared using N-of-1-*pathways* Mahalanobis distance (MD). Presented here is the median absolute MD score for each patient and a given pathway. E.g., The median of the absolute MD scores for the CTC pairs within Patient 6 (ncWnt MD; Pr6) is 0.17. The absolute value is taken because the direction of dysregulation is arbitrary for these pairs. Higher values indicate more heterogeneity for that patient. Also provided is the percentage of pairs found to be significantly dysregulated for a given pathway (FDR < 5%). Again, more dysregulated pairs indicate greater heterogeneity. Within-patient heterogeneity may have clinical value in determining treatment plans. Note that patients Pr3, Pr10, Pr15, and Pr20 are not listed since only one CTC was sequenced for these subjects.

| | Pr6 | Pr17 | Pr9 | Pr11 | Pr14 | Pr19 | Pr18 | Pr22 | Pr21 |
|---|---|---|---|---|---|---|---|---|---|
| Group | Naive | Naive | Naive | Naive | Naive | Resist | Resist | Resist | Resist |
| Num of Pairs | 3 | 6 | 36 | 55 | 55 | 6 | 36 | 45 | 66 |
| ncWnt MD | 0.17 | 0.24 | 0.29 | 0.41 | 0.34 | 0.34 | 0.42 | 0.23 | 0.20 |
| ncWnt % Dys. | 0.00 | 0.00 | 16.67 | 34.55 | 12.73 | 16.67 | 36.11 | 8.89 | 12.12 |
| ErbB2/B3 MD | 0.21 | 0.26 | 0.35 | 0.45 | 0.19 | 0.58 | 0.47 | 0.30 | 0.29 |
| ErbB2/B3 % Dys. | 0.00 | 50.00 | 36.11 | 43.64 | 18.18 | 33.33 | 47.22 | 15.56 | 22.73 |
| SDC4 MD | 0.22 | 0.28 | 0.18 | 0.34 | 0.19 | 0.19 | 0.36 | 0.23 | 0.19 |
| SDC4 % Dys. | 0.00 | 0.00 | 2.78 | 21.82 | 12.73 | 0.00 | 30.56 | 11.11 | 1.52 |
| FOXM1 MD | 0.47 | 0.36 | 0.36 | 0.33 | 0.25 | 0.34 | 0.49 | 0.21 | 0.50 |
| FOXM1 % Dys. | 0.00 | 16.67 | 27.78 | 23.64 | 7.27 | 16.67 | 38.89 | 11.11 | 28.79 |
| S1P1 MD | 0.63 | 0.13 | 0.37 | 0.39 | 0.35 | 0.39 | 0.46 | 0.35 | 0.13 |
| S1P1 % Dys. | 0.00 | 0.00 | 13.89 | 10.91 | 5.45 | 0.00 | 22.22 | 8.89 | 0.00 |

## III. Further investigation of traditionally prioritized pathways

Pathway prioritization by conventional methods can be viewed as a pooling of cells, akin to bulk RNA-seq. Inspection of these identified pathways using our paired analyses highlights the insights possible from single-cell RNA-seq.

**Figure 4** from the main text presents analysis of aggregated within-group pairs quantifying heterogeneity through DEP abundance distances (**Methods Section 2.5**) for 5 pathways derived from N-of-1-*pathways* methods, while **Supplementary Figures S1 and S3** present those distances for pathways derived from WLS and GSEA (**Methods Section 2.3.2**), respectively.

**Figure 5A** from the main text presents analysis of Cell-centric statistics (CCS): Aggregating cell-specific, cross-group pairs to produce single-cell DEPs (**Methods Section 2.4**) for 5 pathways derived from N-of-1-*pathways* methods, while **Supplementary Figures S2 and S4** present those distances for pathways derived from WLS and GSEA (**Methods Section 2.3.2**), respectively.

**a. Weighted least squares (WLS) ranked pathways**

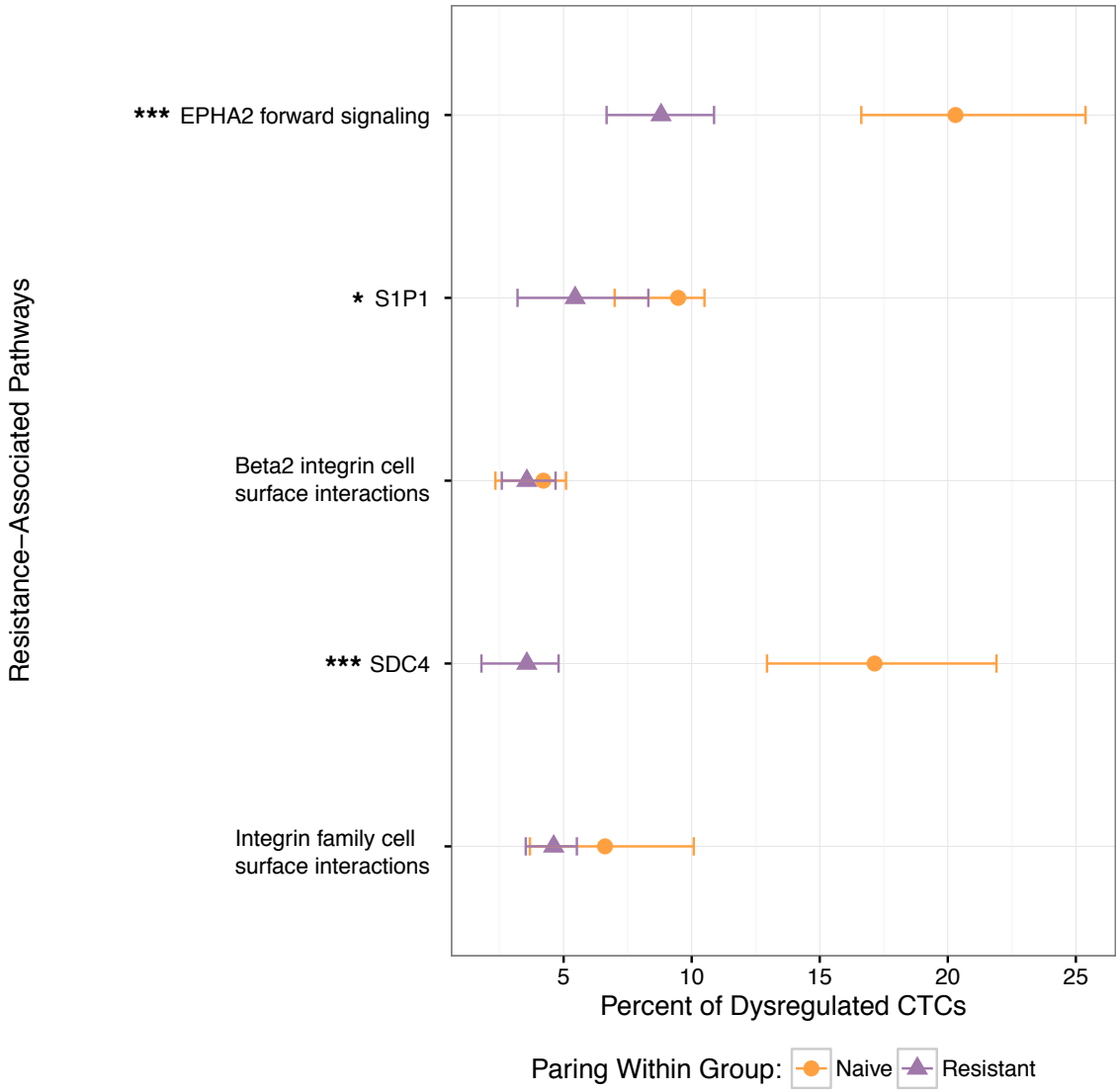


**Fig. S1 Top five resistance-associated pathways ranked by WLS exhibit consistent molecular expression for within-group pairwise CTC comparisons as assessed by aggregated distances.** For the five EZT-resistance pathways ranked by traditional WLS, effect size - as measured by MD scores (Eq. 2) - were calculated for each of the 477 and 635 pairs of circulating tumor cell transcriptomes using combinations within the EZT-resistance group and within the EZT-naïve group, respectively (Fig.1C; Methods 2.4), excluding CTCs paired within patient. For each pathway of a cell pair, a p-value was determined for this MD score (Methods 2.2.1). Each illustrated point represents the proportion of CTC pairs that are significantly differentially expressed for a prioritized pathway (Benjamini-Yekutieli adjusted p-value < 5%). Note that the direction of DEP is arbitrary for within-group pairs by construction. The variability in the statistic is indicated by a 95% bootstrap percentile confidence interval for the proportion of differentially expressed pairs in a given pathway (Methods 2.4.1). In contrast to the proposed method, the CTCs within the EZT-naïve group (N-vs-N) do not exhibit greater heterogeneity than CTCs within the EZT-resistant group (R-vs-R) for every prioritized pathway. ***, **, * indicate p-values smaller than 0.1%, 1%, and 5% respectively, for testing a nonzero difference in DEP prevalence (Methods 2.4.1).

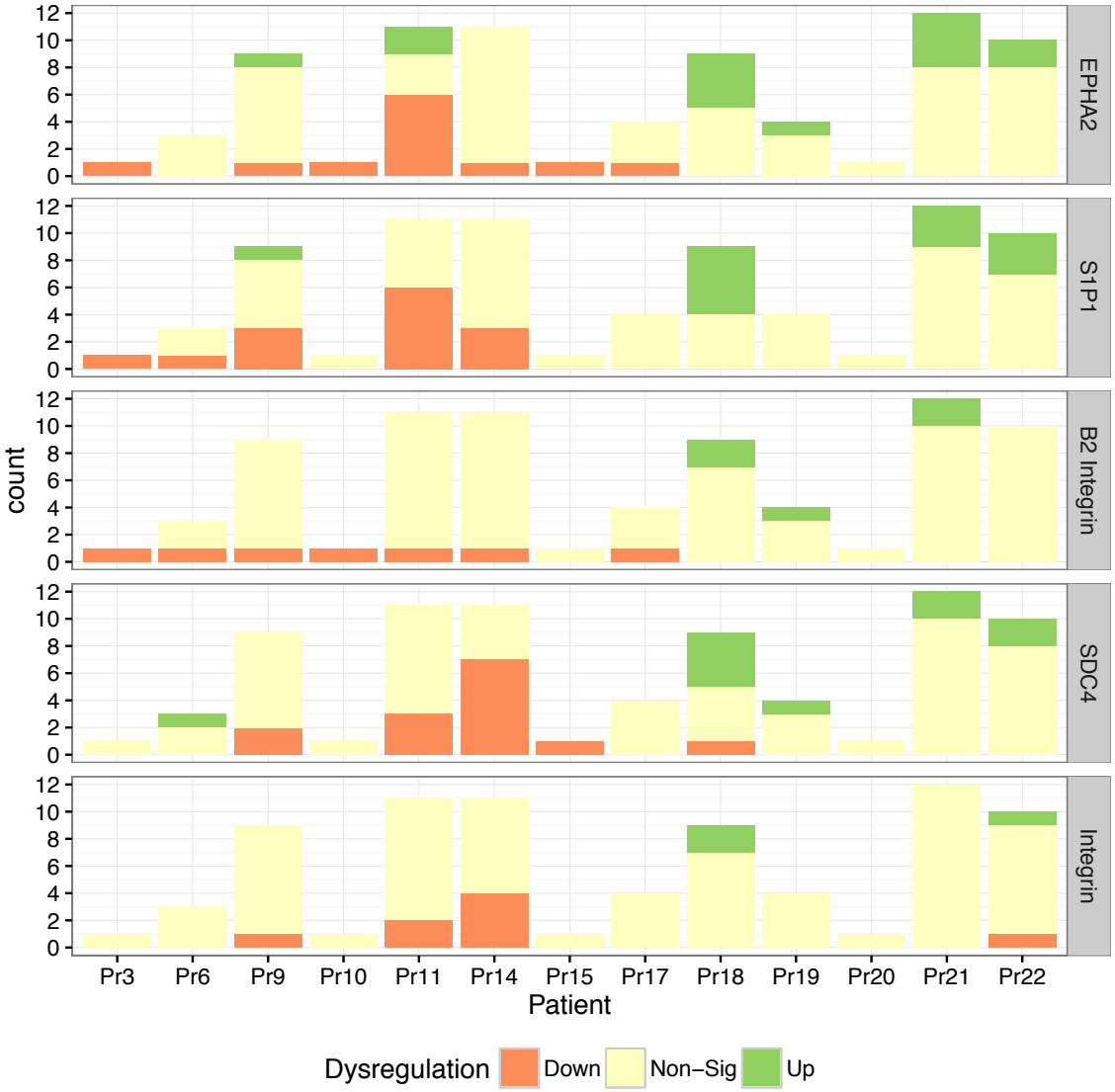


**Fig. S2 Patient-specific transcriptome dynamics of WLS-identified resistance pathways unveiled by cell-centric statistics (CCS) in individual CTCs.** Stacked bar plot of central differential pathway expression. Using MD scores (Eq. 2), the median pathway differential expression effect size of a single CTC was estimated by comparing the pathway mRNAs of this cell of interest to that of all other cross-group CTCs (Fig.1D). For each of the five pathways of a single cell, a 'central DEP status' was determined for the corresponding MD score (Methods 2.5). The majority of significantly DEPs within EZT-naïve CTCs are relatively lower than the resistant CTCs in these five pathways (and, conversely for the resistant CTCs). However, greater heterogeneity in DEPs exists within the EZT-naïve patients. In particular, Pr11 and Pr9 exhibit both up- and down-regulated CTCs. Compared to the identified pathways in Fig. 5, less innate resistance in observed in these pathways. Non-sig = non-significant (pointwise p-value > 5%). Pathway names are to the right of each bar graph.

**a. Gene set enrichment analysis (GSEA) ranked pathways**

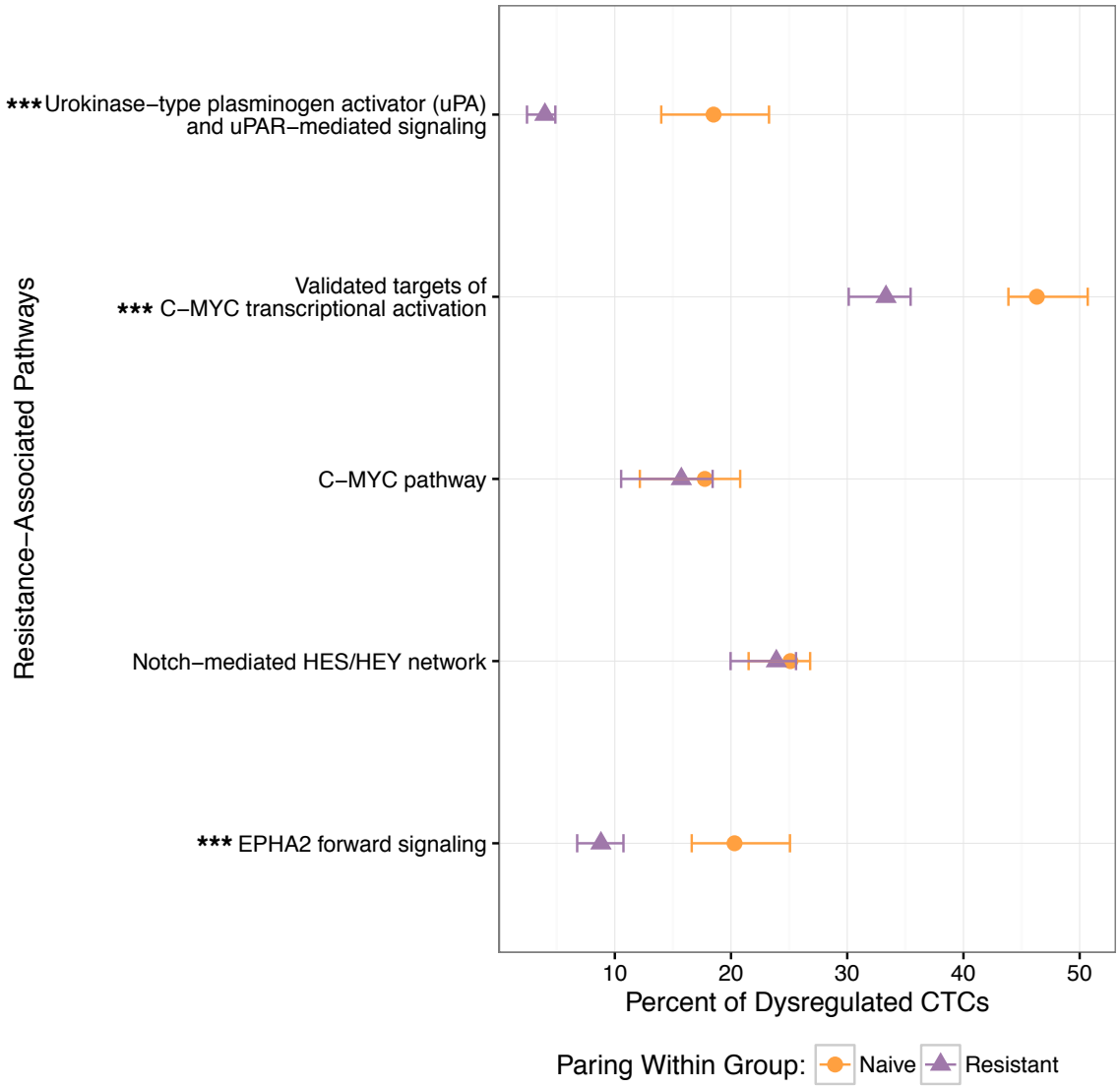


**Fig. S3 Top five resistance-associated pathways ranked by GSEA do not exhibit differential consistency molecular expression for within-group pairwise CTC comparisons as assessed by aggregated distances.** For the five EZT-resistance pathways ranked by traditional WLS, effect size - as measured by MD scores (Eq. 2) - were calculated for each of the 477 and 635 pairs of circulating tumor cell transcriptomes using combinations within the EZT-resistance group and within the EZT-naïve group, respectively (Fig.1C; Methods 2.4), excluding CTCs paired within patient. For each pathway of a cell pair, a p-value was determined for this MD score (Methods 2.2.1). Each illustrated point represents the proportion of CTC pairs that are significantly differentially expressed for a prioritized pathway (Benjamini-Yekutieli adjusted p-value < 5%). Note that the direction of DEP is arbitrary for within-group pairs by construction. The variability in the statistic is indicated by a 95% bootstrap percentile confidence interval for the proportion of differentially expressed pairs in a given pathway (Methods 2.4.1). In contrast to the proposed method, the CTCs within the EZT-naïve group (N-vs-N) do not exhibit greater heterogeneity than CTCs within the EZT-resistant group (R-vs-R) for every prioritized pathway. Additionally, *Validated targets of C-MYC* demonstrates very high rates of dysregulation for both groups. ***, **, * indicate p-values smaller than 0.1%, 1%, and 5% respectively, for testing a nonzero difference in DEP prevalence (Methods 2.4.1).

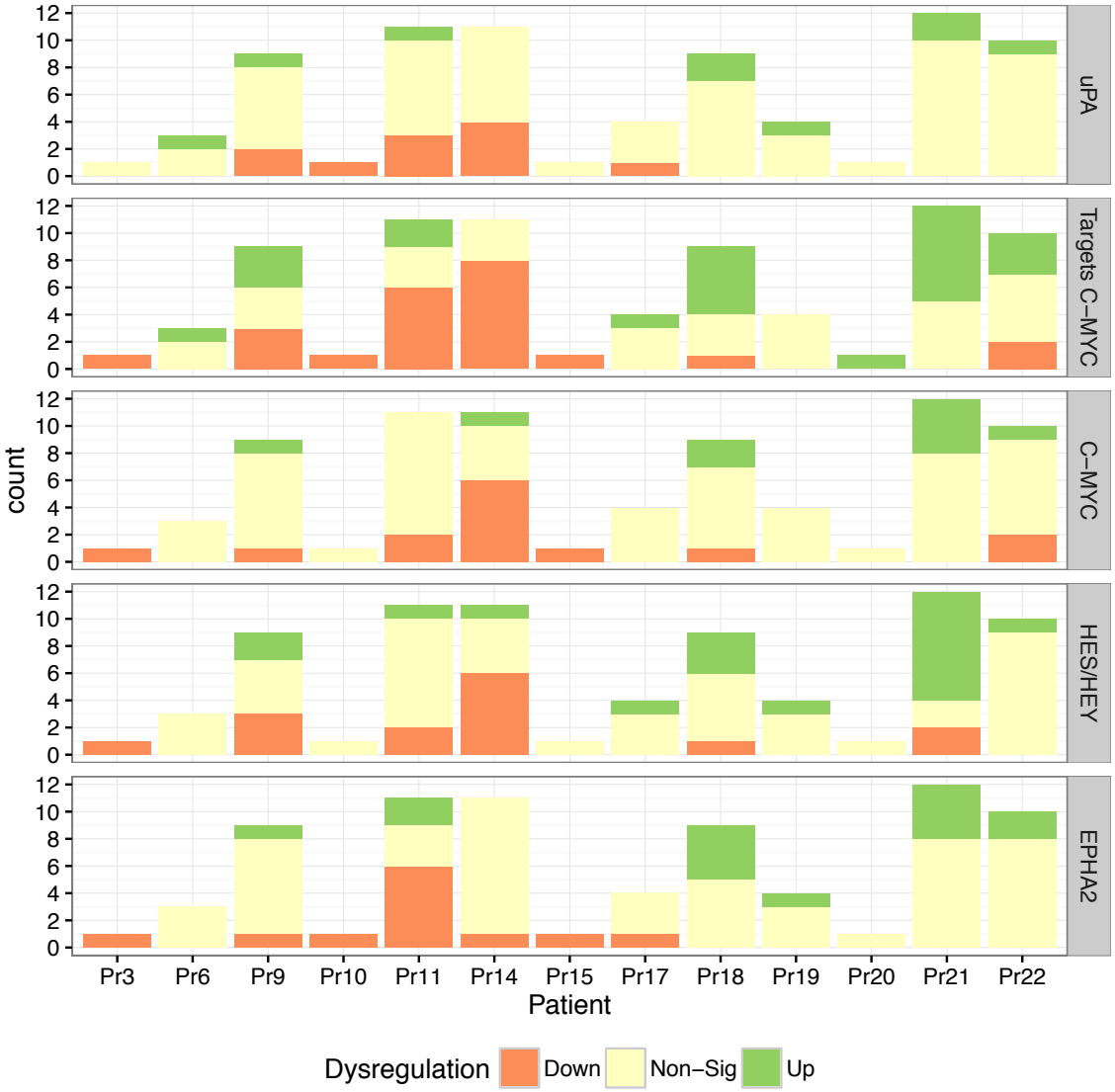


**Fig. S4 Patient-specific transcriptome dynamics of GSEA-identified resistance pathways unveiled by cell-centric statistics (CCS) in individual CTCs.** Stacked bar plot of central differential pathway expression. Using MD scores (Eq. 2), the median pathway differential expression effect size of a single CTC was estimated by comparing the pathway mRNAs of this cell of interest to that of all other cross-group CTCs (Fig.1D). For each of the five pathways of a single cell, a 'central DEP status' was determined for the corresponding MD score (Methods 2.5). The majority of significantly DEPs within EZT-naïve CTCs are relatively lower than the resistant CTCs in these five pathways (and, conversely for the resistant CTCs). In contrast to the identified pathways in Fig. 5, heterogeneity is evident in both treatment groups. In particular, Pr11, Pr21, Pr22, and Pr9 exhibit both up- and down-regulated CTCs. Innate resistance in observed for Pr9, Pr11, and Pr14. Non-sig = non-significant (pointwise p-value > 5%). Pathway names are to the right of each bar graph.