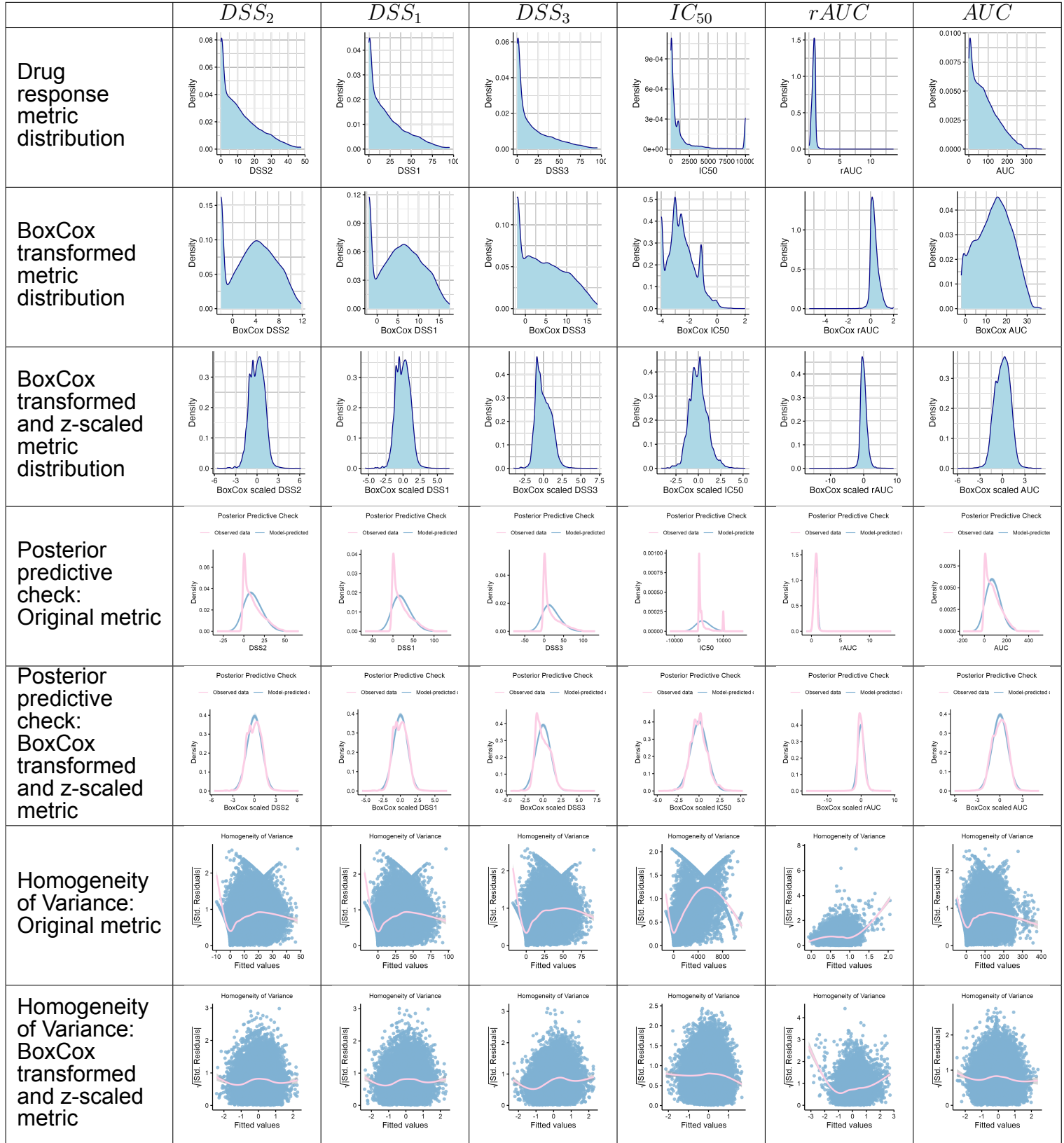


Suppl. Table 1. The influence of BoxCox transformation and z-scaling on the six different response metrics (DSS_2 , DSS_1 , DSS_3 , IC_{50} , $rAUC$, AUC).



The first 3 rows show the distribution of each metric with their original scale (first row), and after BoxCox transformation (second row), and after BoxCox transformation and z-scaling (third row). The last four rows show diagnostics plots of the medium-specific LMM (used as an example here) before and after BoxCox transformation and z-scaling the different response metrics. The fourth row shows plots to check the homogeneity of variance for the medium-specific model fitted on each response metric before data transformation. The fifth row shows plots to check the homogeneity of variance for the medium-specific model fitted on BoxCox transformed and z-scaled metrics, demonstrating that homogeneity of variance improves post transformation. The sixth row shows posterior predictive check of the medium-specific model fitted on each metric prior to data transformation. The seventh row shows posterior predictive check for medium-specific model fitted on BoxCox transformed and z-scaled metrics, indicating that the model fitted on metrics post transformation better captures the structure of the drug response data.

Suppl. Table 2. Model comparison between different random effect structures when modeling experimental variables, where AIC values indicate model fit for the response data taking into account the model complexity.

Model	AIC	p-value
DSS2 Box-Cox scaled Time until usage + (1 Drug)	109166	
DSS2 Box-Cox scaled Time until usage usage + (1 Patient)	102509	
DSS2 Box-Cox scaled Time until usage + (1 Patient) + (1 Drug)	102437	<0.001
DSS2 Box-Cox scaled Time until usage + (1 + Time until usage Drug)	106187	1
DSS2 Box-Cox scaled Time until usage + (1 Patient) + (1 + Time until usage Drug)	98642	<0.001
DSS2 Box-Cox scaled Time until usage usage + (1 Patient) + (0 + Time until usage Drug)	98642	
DSS2 Box-Cox scaled Time until usage + (1 Patient) + (1 Drug) + (0 + Time until usage Drug)	98644	1.0

The p-values indicate whether the model fits significantly better than the one above. The p-values shown for nested model comparisons; non-nested models lack a valid basis for statistical comparison. The model in bold is the selected model used in the presented analysis, where time until sample usage was used as an example.

Suppl. Table 3. Model comparison between different random effect structures when modeling biospecimen variables, where AIC values indicate model fit for the response data taking into account the model complexity.

Model	AIC	p-value
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Patient) + (1 Drug) + (1 Lab)	96044	
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Patient) + (1 Drug) + (1 Sample) + (1 Tissue) + (1 Disease status)	96371	1.0
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Drug) + (1 Sample) + (1 Tissue) + (1 Disease status) + (1 Lab)	100223	
DSS2 BoxCox scaled Sample + Tissue + Disease status + (1 Patient) + (0 + Lab Drug)	87556	<0.001
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Patient) + (Lab Drug)	87556	
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Patient) + (Sample Drug) + (Tissue Drug) + (Disease status Drug)	94081	<0.001
DSS2 Box-Cox scaled Sample + Tissue + Disease status + (1 Patient) + (Sample Drug) + (Tissue Drug) + (Disease status Drug) + (1 Lab)	94379	1

The p-value indicates whether the model fits significantly better than the one above. The p-values are shown for nested model comparisons; non-nested models lack a valid basis for statistical comparison. The biospecimen model in bold is the final model used in this analysis.

Suppl. Table 4. Comparison of clinical and molecular characteristics of patients between the cohorts.

Characteristic	Study				p-value ¹	q-value ²
	Beat AML N=631	Helsinki N=97	Karolinska N=192	Oslo N=78		
Sex					0.065	0.095
Female	284 (45%)	55 (57%)	78 (43%)	42 (54%)		
Male	346 (55%)	42 (43%)	104 (57%)	36 (46%)		
Missing	1 (0.2%)	0 (0%)	10 (5.2%)	0 (0%)		
Age, y	61 (46–70)	64 (53–70)	66 (53–74)	61 (50–65)	<0.001	<0.001
Missing	21 (3.3%)	0 (0%)	11 (5.7%)	0 (0%)		
AML type					<0.001	<0.001
Diagnosis	397 (66%)	58 (60%)	173 (95%)	69 (88%)		
Refractory	138 (23%)	8 (8.2%)	0 (0%)	0 (0%)		
Relapse	45 (7.5%)	31 (32%)	10 (5.5%)	8 (10%)		
Remission	19 (3.2%)	0 (0%)	0 (0%)	1 (1.3%)		
Missing	32 (5.1%)	0 (0%)	9 (4.7%)	0 (0%)		
FAB					<0.001	<0.001
Other/NOS	7 (2.2%)	14 (19%)	0 (NA%)	6 (8.1%)		
M0, M1 or M2	159 (50%)	40 (56%)	0 (NA%)	35 (47%)		
M3	15 (4.7%)	0 (0%)	0 (NA%)	3 (4.1%)		
M4 or M5	132 (42%)	18 (25%)	0 (NA%)	26 (35%)		
M6	2 (0.6%)	0 (0%)	0 (NA%)	3 (4.1%)		
M7	2 (0.6%)	0 (0%)	0 (NA%)	1 (1.4%)		
Missing	314 (50%)	25 (26%)	192 (100%)	4 (5.1%)		
Cytogenetic alterations					<0.001	<0.001
Normal karyotype	367 (58%)	67 (69%)	82 (43%)	59 (76%)		
Other / Unknown	97 (15%)	11 (11%)	77 (40%)	9 (12%)		
Adverse risk karyotype	167 (26%)	19 (20%)	33 (17%)	10 (13%)		
Complex karyotype	159 (95%)	12 (63%)	23 (70%)	8 (80%)		
del(5)/del(5q)	1 (0.6%)	4 (21%)	2 (6.1%)	0 (0%)		
del(7)/del(7q)	7 (4.2%)	3 (16%)	8 (24%)	2 (20%)		
Previous transplant	192 (30%)	26 (31%)	40 (30%)	37 (47%)	0.023	0.037
Missing	0 (0%)	13 (13%)	60 (31%)	0 (0%)		
Blasts percentage	70 (38–85)	60 (35–70)	50 (29–70)	NA (NA–NA)	<0.001	<0.001
Missing	163 (26%)	17 (18%)	23 (12%)	78 (100%)		
ELN 2017/2022					<0.001	<0.001
Adverse	148 (39%)	20 (23%)	62 (40%)	22 (28%)		
Favorable	150 (39%)	18 (21%)	49 (31%)	17 (22%)		
Intermediate	85 (22%)	49 (56%)	45 (29%)	39 (50%)		
Missing	248 (39%)	10 (10%)	36 (19%)	0 (0%)		
Somatic mutations					0.009	0.017
NPM1	148 (23%)	22 (23%)	66 (36%)	20 (26%)	0.009	0.017
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	0 (0%)		
FLT3-ITD/TKD	68 (11%)	6 (6.3%)	67 (35%)	23 (29%)	<0.001	<0.001
Missing	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)		
SRSF2	78 (12%)	8 (8.4%)	0 (NA%)	2 (2.6%)	0.013	0.023
Missing	0 (0%)	2 (2.1%)	192 (100%)	0 (0%)		
RUNX1	71 (11%)	10 (11%)	19 (10%)	5 (6.4%)	0.6	0.7
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	0 (0%)		
TP53	55 (8.7%)	3 (3.2%)	12 (6.5%)	0 (NA%)	0.13	0.2
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	78 (100%)		
IDH1	45 (7.1%)	11 (12%)	8 (4.3%)	0 (NA%)	0.078	0.11
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	78 (100%)		
IDH2	70 (11%)	22 (23%)	34 (18%)	0 (NA%)	0.001	0.003
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	78 (100%)		
CEBPA	37 (5.9%)	8 (8.4%)	0 (NA%)	7 (9.0%)	0.3	0.4
Missing	0 (0%)	2 (2.1%)	192 (100%)	0 (0%)		
ASXL1	61 (9.7%)	13 (14%)	18 (9.7%)	7 (9.0%)	0.7	0.7
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	0 (0%)		
STAG2	37 (5.9%)	7 (7.4%)	11 (5.9%)	2 (2.6%)	0.6	0.7

Characteristic	Study				p-value ¹	q-value ²
	Beat AML N=631	Helsinki N=97	Karolinska N=192	Oslo N=78		
Missing	0 (0%)	2 (2.1%)	7 (3.6%)	0 (0%)		
KRAS or NRAS	125 (20%)	11 (12%)	28 (15%)	5 (6.4%)	0.006	0.012
Missing	0 (0%)	2 (2.1%)	0 (0%)	0 (0%)		

¹Pearson's Chi-squared test; Kruskal-Wallis rank sum test; Fisher's exact test. ²False discovery rate correction for multiple testing.

Suppl. Table 5. Comparison of model estimates obtained from the original model (LMM without genetics) and after inclusion of RAS, FLT3, and NPM1 mutation status as covariates (LMM with genetics). Change in Coef represents the difference in coefficients, calculated as the estimate from the model including genetic covariates minus the estimate from the original model.

Experimental variable		Model without genetics		Model with genetics		Change in	
		Coef.	Adj. p-value	Coef.	Adj. p-value	Coef.	Adj. p-value
Medium	MCM	0.112	1	0.107	1	-0.005	None*
	RPMI + FBS	0.695	<0.001	0.732	<0.001	0.037	<0.0001
	RAS	-	-	-0.132	<0.001	-	-
	FLT3	-	-	0.102	0.031	-	-
	NPM1	-	-	0.142	<0.001	-	-
Positive control + Doses + Readout + Cell Counting	BzCl + 5 + CellTiter-Glo + Trypan Blue	-0.676	<0.001	-0.714	<0.001	-0.038	<0.0001
	RAS	-	-	-0.136	0.03	-	-
	FLT3	-	-	0.104	<0.001	-	-
	NPM1	-	-	0.140	<0.001	-	-
Time until sample usage	>2h	0.473	<0.001	0.518	<0.001	0.045	<0.0001
	RAS	-	-	-0.071	0.500	-	-
	FLT3	-	-	0.065	0.825	-	-
	NPM1	-	-	0.145	<0.001	-	-
Centrifugation Procedure	LymphoPrep™ gradient	-0.503	<0.001	-0.562	<0.001	-0.059	<0.0001
	RAS	-	-	-0.102	0.040	-	-
	FLT3	-	-	0.115	0.056	-	-
	NPM1	-	-	0.137	<0.001	-	-
Plate Reader	EnSight	-0.643	<0.001	-0.701	<0.001	-0.058	<0.0001
	EnVision	-0.630	<0.001	-0.675	<0.001	-0.045	<0.0001
	Pherastar or Cytation	-0.777	<0.001	-0.786	<0.001	-0.009	<0.0001
	VictorX	-0.473	0.002	-0.519	<0.001	-0.046	0.0002
	RAS	-	-	-0.133	<0.001	-	-
	FLT3	-	-	0.089	0.103	-	-
	NPM1	-	-	0.141	<0.001	-	-
Cells	5000	-0.643	<0.001	-0.664	<0.001	-0.021	<0.0001
	RAS	-	-	-0.113	0.010	-	-
	FLT3	-	-	0.045	1	-	-
	NPM1	-	-	0.138	<0.001	-	-

*Adjusted p-values exceeding 1 were both truncated to 1.

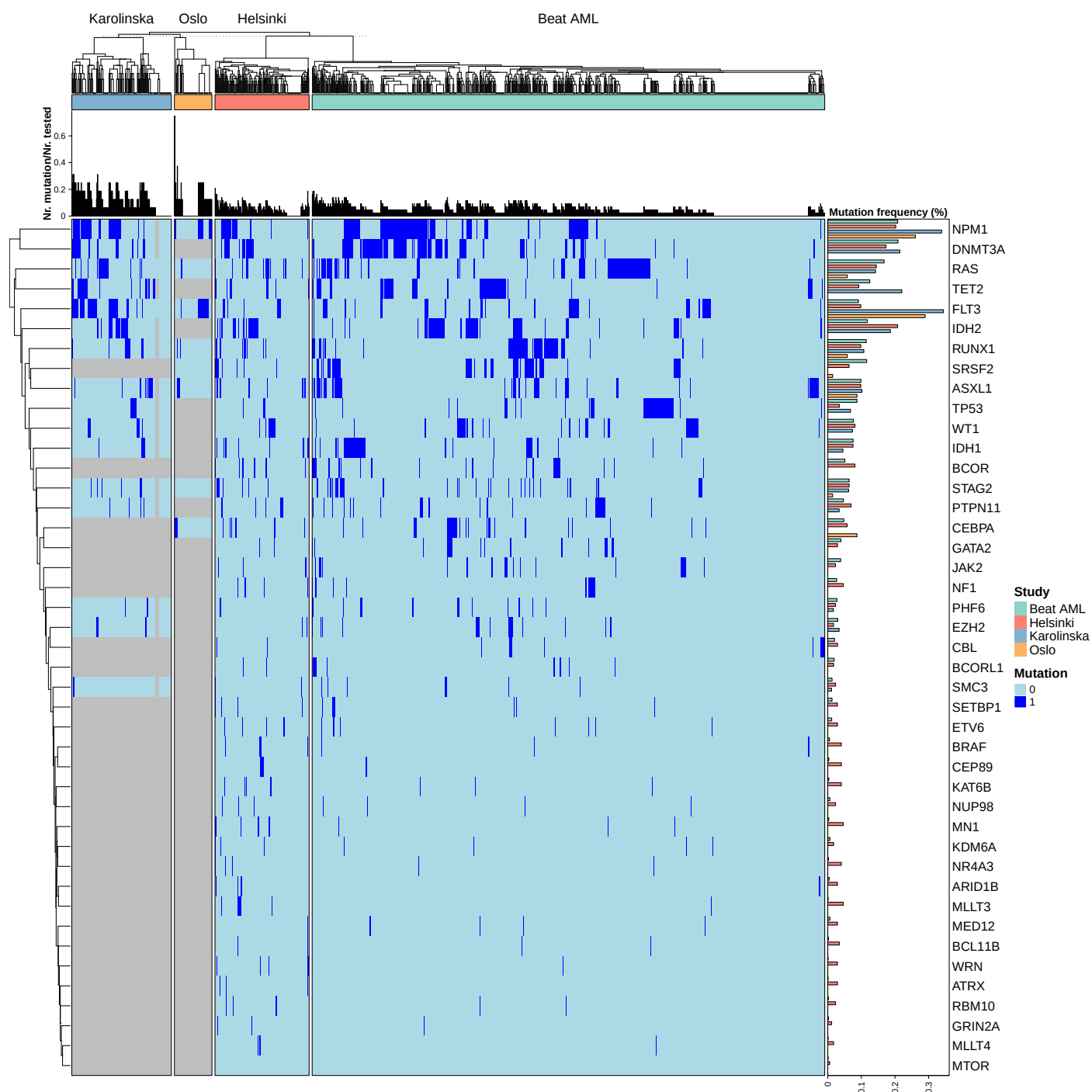
Suppl. Table 6. Comparison of medium model estimates obtained from original LMM including all samples from all patients and after stratification of specific characteristics. Coefficients are shown for the substratified models. Change in Coef represents the difference in coefficients, calculated as the estimate from the substratified model based on each characteristic minus the estimate from the original model.

Substratification Category		Medium	Substratified Model		Change	
			Coef.	Adj. p-value	Coef.	Adj. p-value
ELN	Favorable	MCM	0.212	0.822	-0.005	0.178
		RPMI + FBS	0.721	<0.001	0.100	<0.0001
	Intermediate	MCM	0.118	1	-0.038	None*
		RPMI + FBS	0.729	<0.001	0.034	<0.0001
	Adverse	MCM	0.007	1	-0.105	None*
		RPMI + FBS	0.607	<0.001	-0.088	<0.0001
Cytogenetic Alteration	Normal Karyotype	MCM	0.130	1	0.018	None*
		RPMI + FBS	0.688	<0.001	-0.007	<0.0001
	Adverse risk karyotype	MCM	0.161	1	0.049	None*
		RPMI + FBS	0.690	<0.001	-0.005	<0.0001
AML Type	Diagnosis	MCM	0.069	1	-0.043	None*
		RPMI + FBS	0.681	<0.001	-0.014	<0.0001
	Relapse	MCM	0.371	0.1	0.259	0.9
		RPMI + FBS	0.802	<0.001	0.107	<0.0001

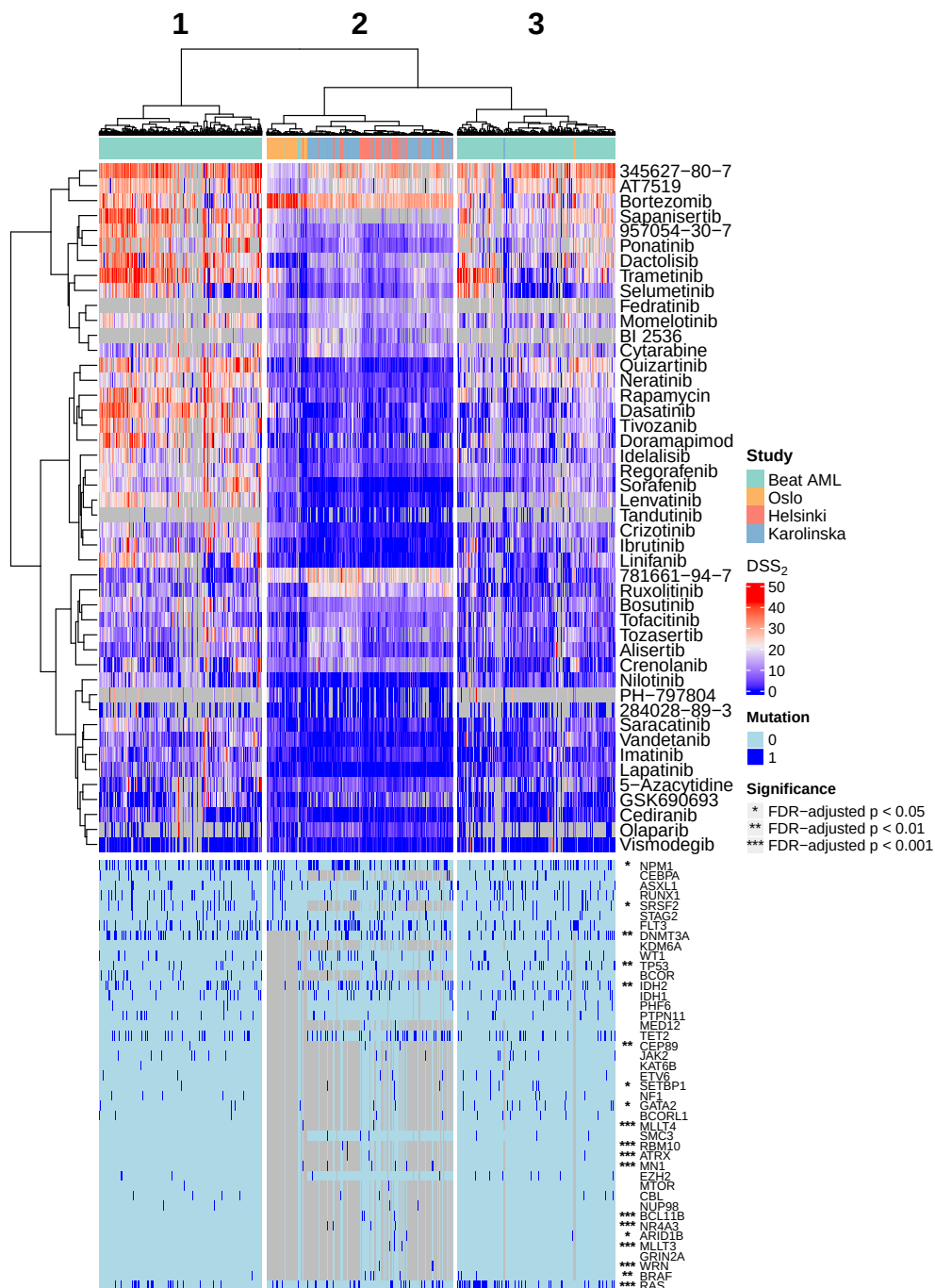
*Adjusted p-values exceeding 1 were both truncated to 1.

Suppl. Table 7. Summary of transformed risk ratios and corresponding E-values. Risk ratios were obtained by transforming the fixed-effect coefficients from the LMMs to the RR scale. E-values were calculated for each association to estimate the minimum strength of association that an unmeasured confounder would need to have with both the exposure (experimental variable) and the outcome (DSS_2) to fully explain the observed effect. Larger E-values indicate that a stronger unmeasured confounder would be required to negate the observed association.

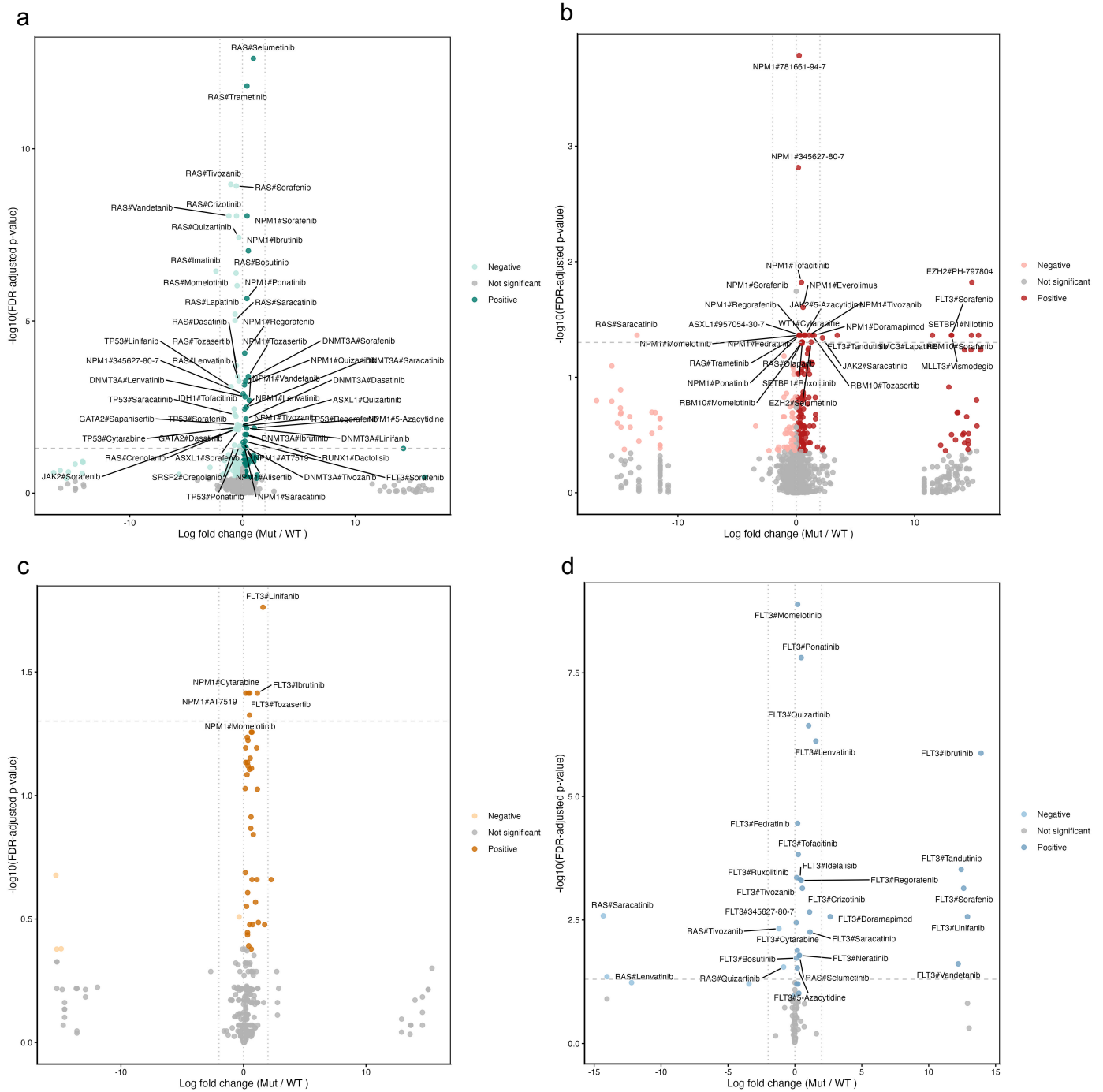
Experimental variable		Risk Ratio	E-value
Time until sample usage	<2h	1.539	2.449
Medium	MCM	1.109	1.456
	RPMI + FBS	1.871	3.147
Cells	5000	0.535	3.143
Positive control + Doses + Readout + Cell Counting	BzCl + 5 + CellTiter-Glo + Trypan Blue	0.544	3.077
Centrifugation Procedure	LymphoPrep™ gradient	0.633	2.538
Plate reader	Enight	0.557	2.990
	EnVision	0.564	2.945
	Pherastar or Cytation	0.493	3.472
	VictorX	0.650	2.447



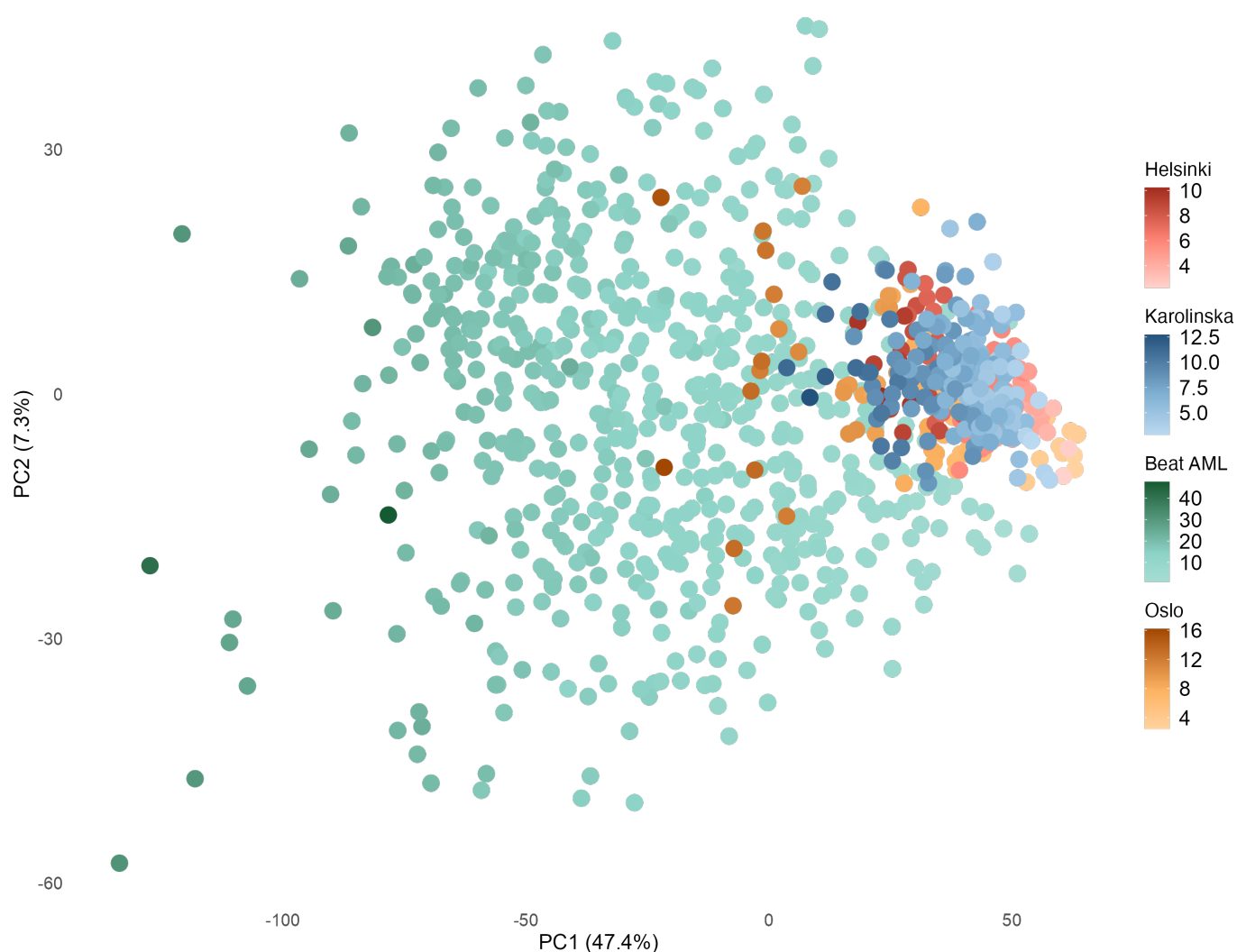
Suppl. Fig. 1. Gene mutation frequency between the cohorts. Grey color indicates that the gene was not included in the mutation panel of the particular study.



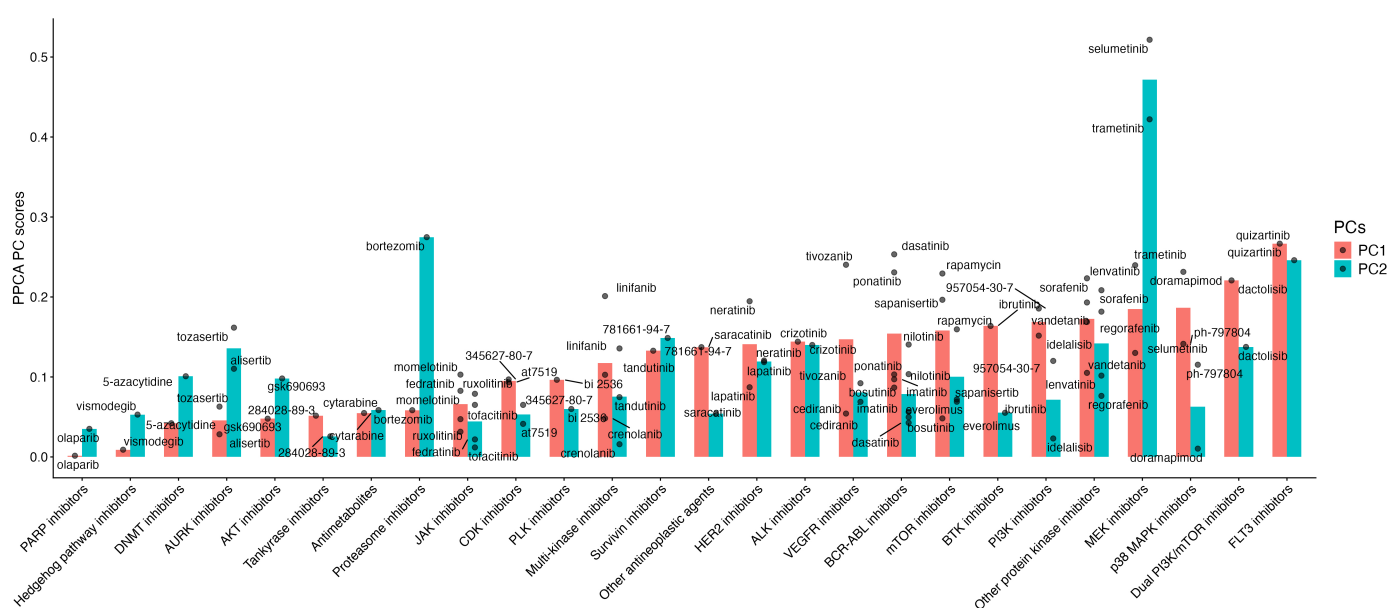
Suppl. Fig. 2. Drug response clustering and gene mutation enrichment analysis. Upper panel, hierarchical clustering of the patients using DSS_2 responses and Ward D2 algorithm. Bottom panel, mutation profiles across the clusters, where grey color indicates that the gene was not included in the mutation panel of the particular study. Enrichment of mutations in the clusters was tested with Fisher's exact test.



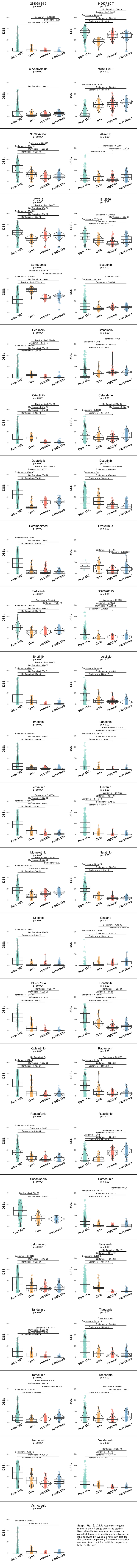
Suppl. Fig. 3. Statistical associations between DSS_2 and gene mutations. (a) Beat AML, (b) Helsinki, (c) Oslo, and (d) Karolinska cohort. Negative associations mean that patients with the mutation show lower DSS_2 compared to wild type, and positive associations mean that patients with the mutation show higher DSS_2 compared to wild type. When using the cutoff of log-fold-change greater than 2 (vertical dotted lines), no common associations remained between the cohorts. The horizontal dotted lines indicate FDR-adjusted cutoff of $p = 0.05$.

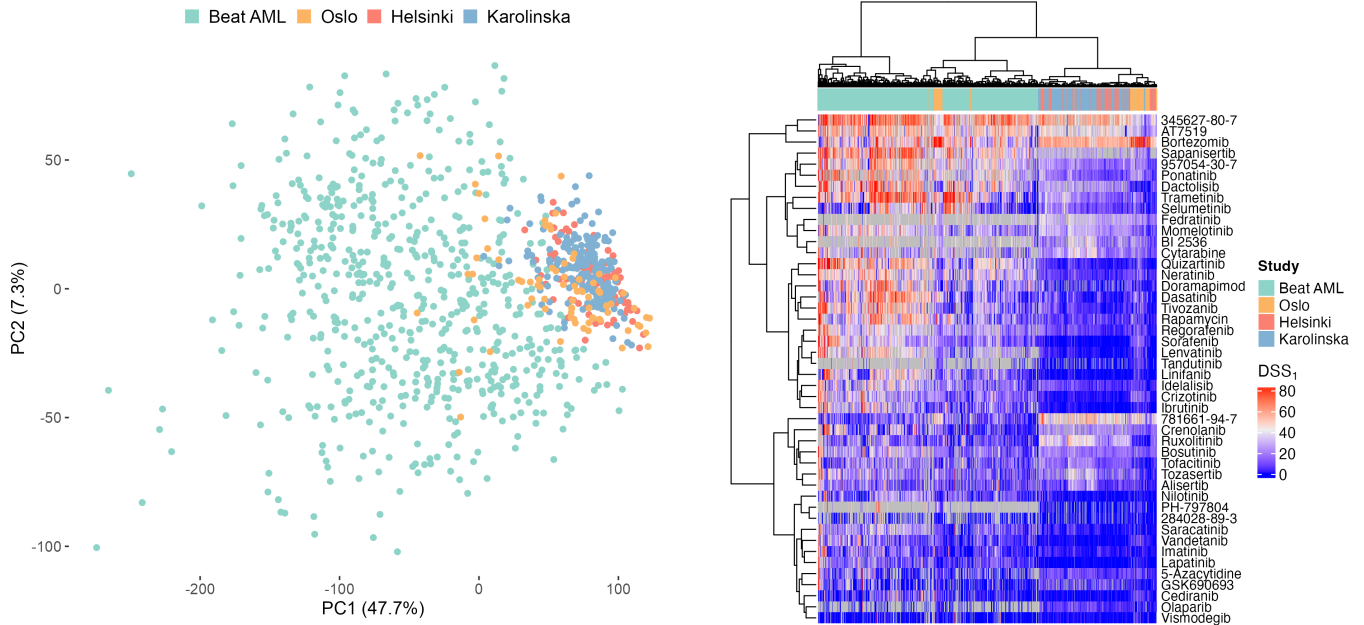


Suppl. Fig. 4. PPCA was used to visualise the DSS_2 profiles from the four studies. Each point is a patient sample. The color scale is based on the average DSS_2 score across the 46 common drugs for each sample, with different color scales for each lab.

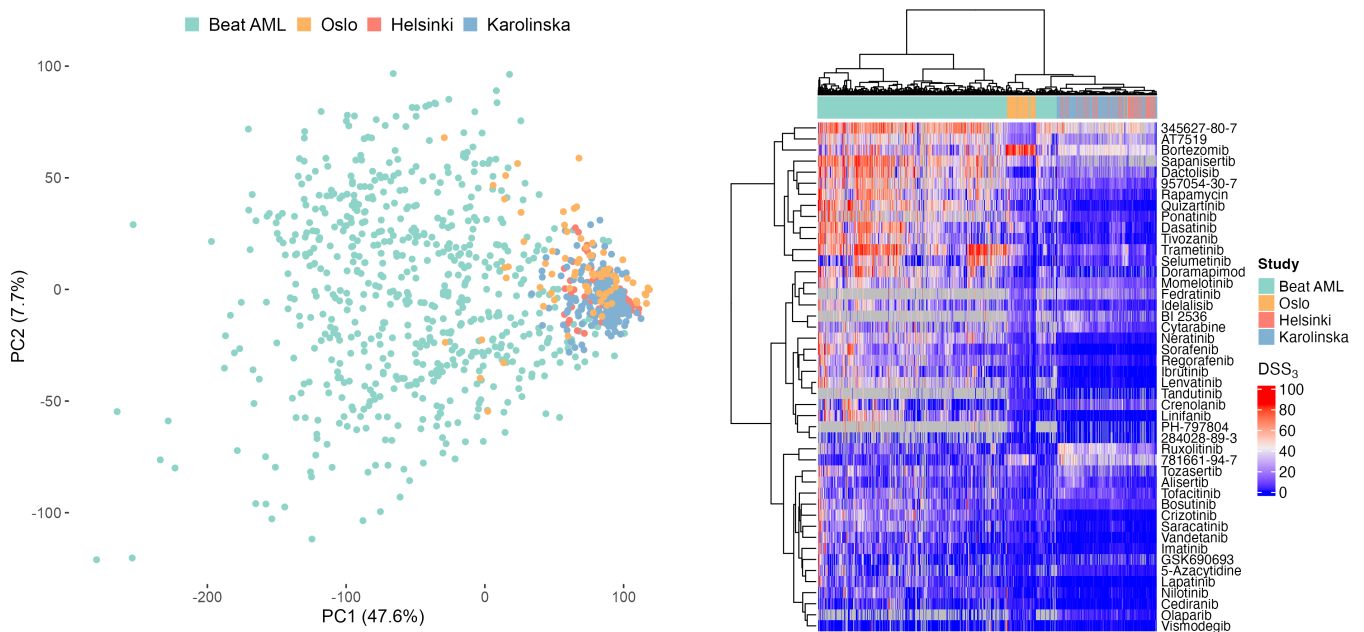


Suppl. Fig. 5. Contributions of drugs and drug class to PC1 and PC2 loadings in the PPCA plot of Fig.1d. Points correspond to individual drugs (46 common compounds), and the bar heights indicate the average PC loadings in each class.

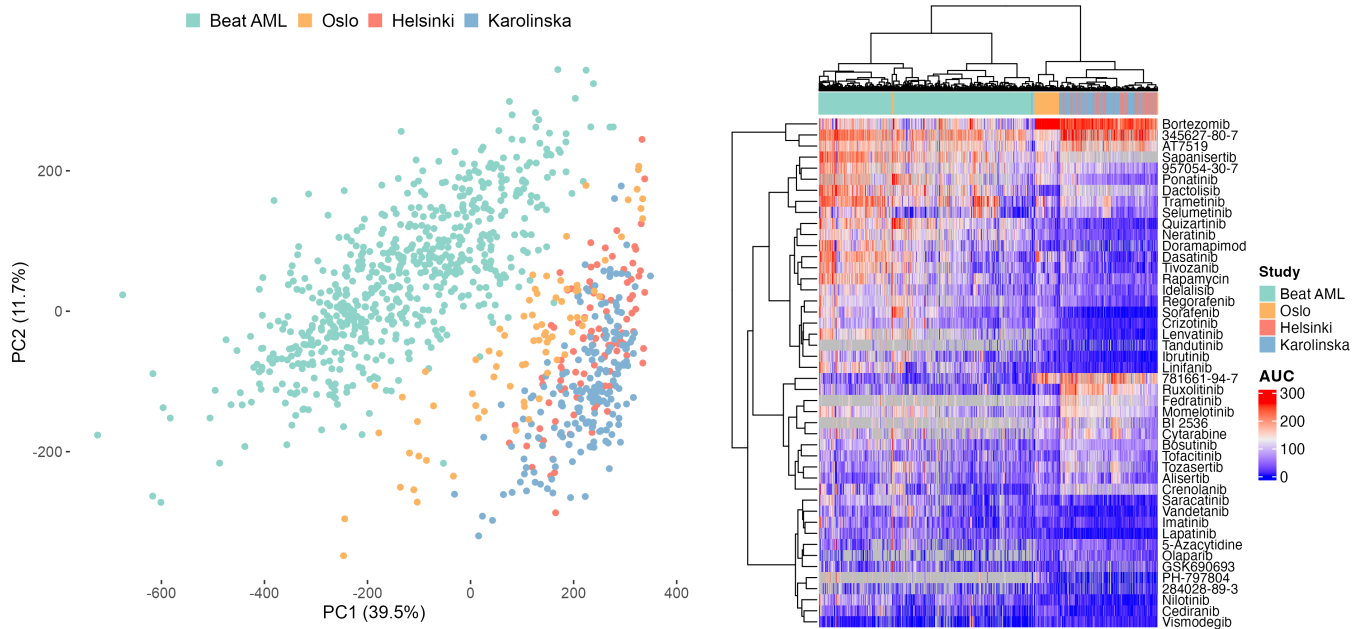




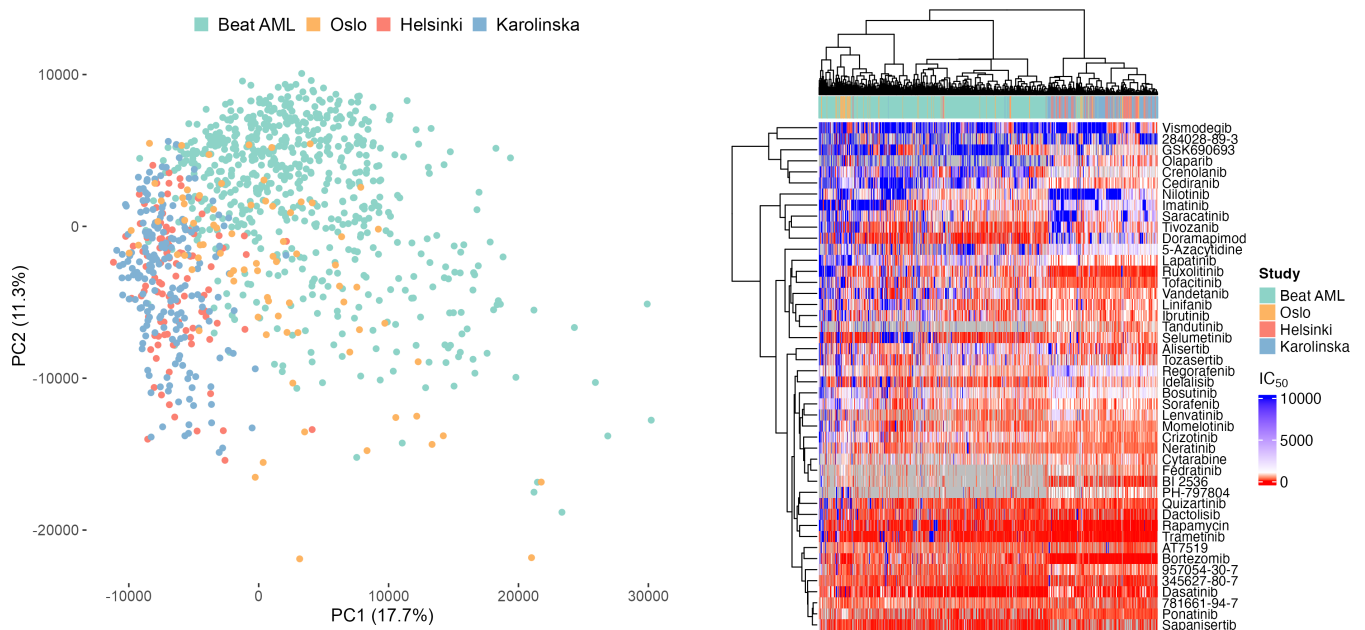
Suppl. Fig. 7. *Left:* PPCA was used to visualise the DSS_1 profiles of the patients from the four studies. *Right:* Heatmap of DSS_1 responses across the four different labs (color bar at the top), where the columns correspond to the patient samples and rows to the 46 overlapping drugs. The high DSS_1 levels indicate high drug sensitivity. The grey color indicates missing data, i.e., a patient sample had not been tested for a particular drug.



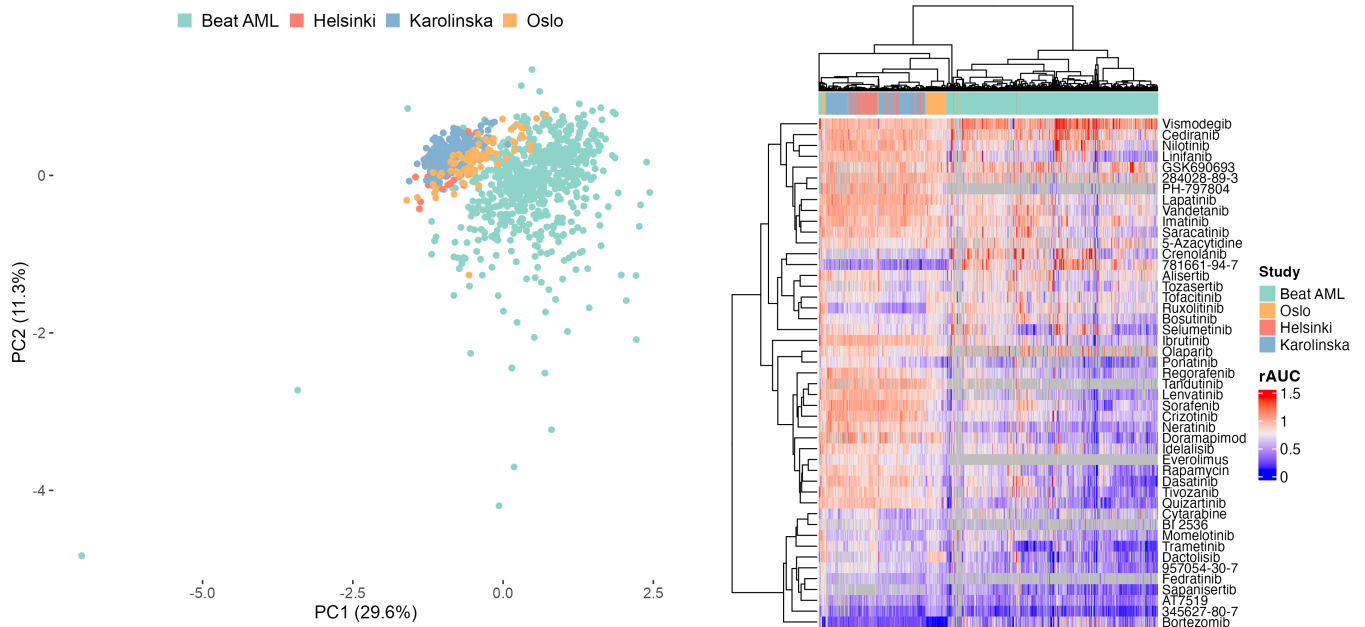
Suppl. Fig. 8. *Left:* PPCA was used to visualise the DSS_3 profiles of the patients from the four studies. *Right:* Heatmap of DSS_3 responses across the four different labs (color bar at the top), where the columns correspond to the patient samples and rows to the 46 overlapping drugs. The high DSS_3 levels indicate high drug sensitivity. The grey color indicates missing data, i.e., a patient sample had not been tested for a particular drug.



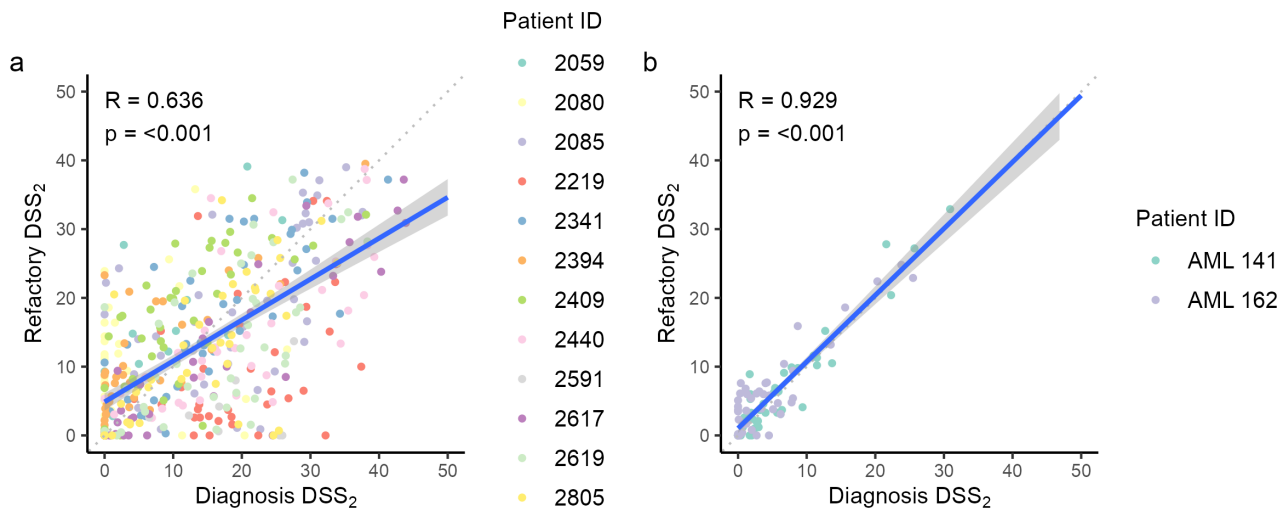
Suppl. Fig. 9. *Left:* PPCA was used to visualise the AUC profiles of the patients from the four studies. *Right:* Heatmap of AUC responses across the four different labs (color bar at the top), where the columns correspond to the patient samples and rows to the 46 overlapping drugs. The high AUC levels indicate high overall drug activity. The grey color indicates missing data, i.e., a patient sample had not been tested for a particular drug.



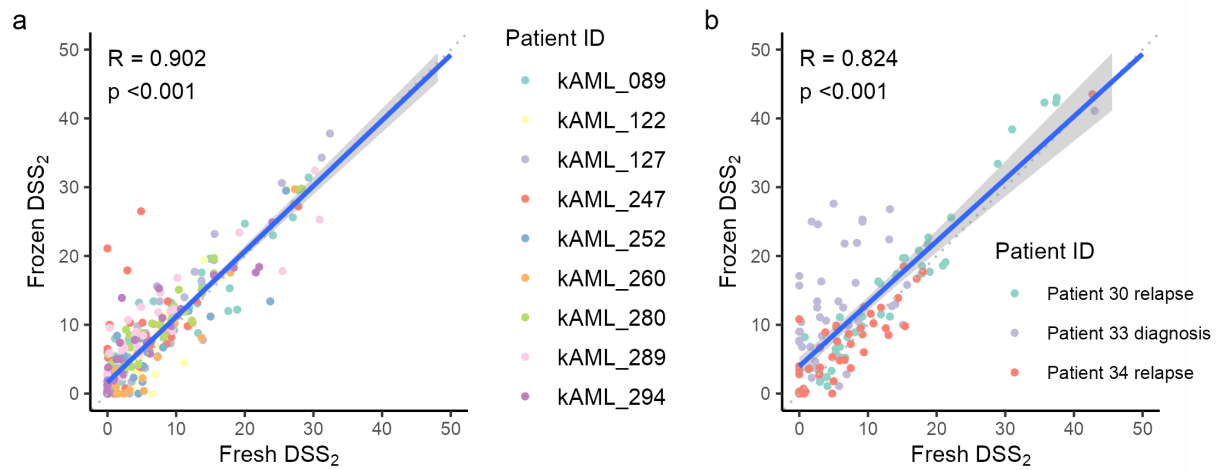
Suppl. Fig. 10. *Left:* PPCA was used to visualise the IC_{50} profiles of the patients from the four studies. *Right:* Heatmap of IC_{50} responses across the four different labs (color bar at the top), where the columns correspond to the patient samples and rows to the 46 overlapping drugs. The low IC_{50} levels indicate low potency. The grey color indicates missing data, i.e., a patient sample had not been tested for a particular drug.



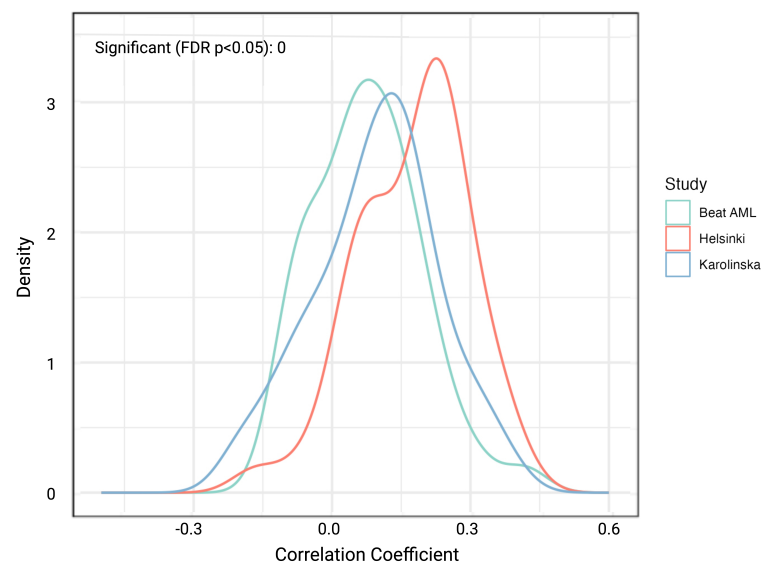
Suppl. Fig. 11. Left: PPCA was used to visualise the $rAUC$ profiles of the patients from the four studies. Right: Heatmap of $rAUC$ responses across the four different labs (color bar at the top), where the columns correspond to the patient samples and rows to the 46 overlapping drugs. The high $rAUC$ levels indicate high overall drug activity. The grey color indicates missing data, i.e., a patient sample had not been tested for a particular drug.



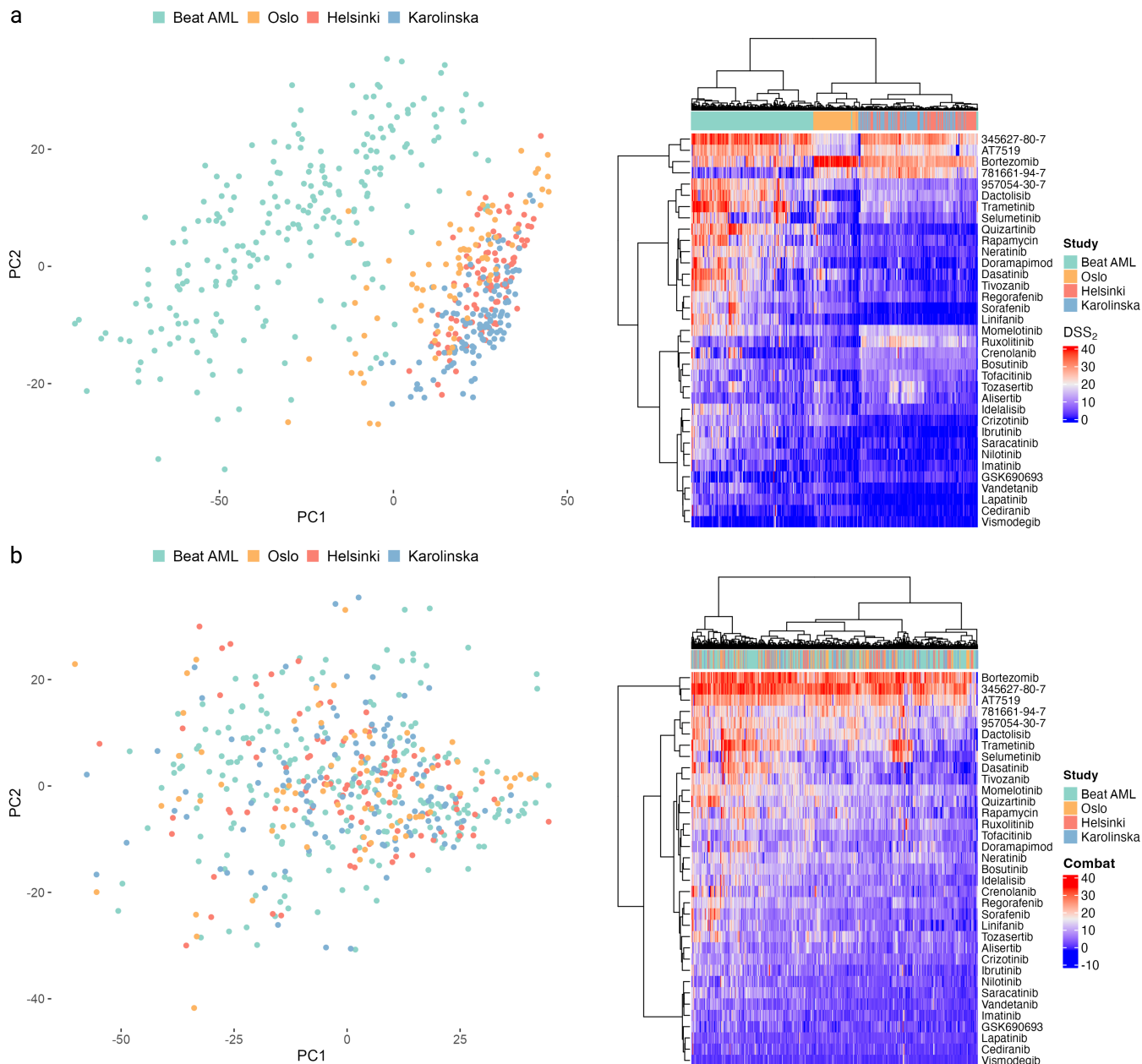
Suppl. Fig. 12. a-b Pearson correlation of DSS_2 responses between paired Diagnosis and Refractory samples from the same patients in (a) Beat AML ($n=12$) and (b) Helsinki cohorts ($n=2$) across the 46 common drugs (the points with the same color). The correlation line (blue solid line) is below the diagonal (grey dotted line), indicating that the refractory samples lead generally to lower drug responses compared to diagnostic samples.



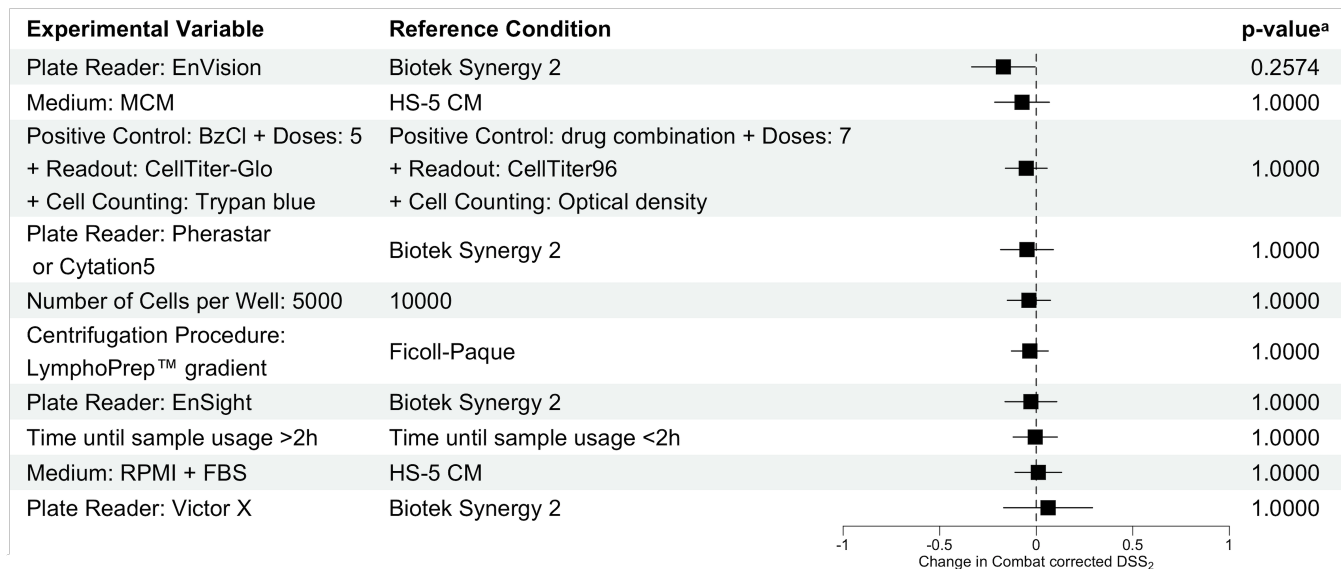
Suppl. Fig. 13. a-b Pearson correlation of DSS_2 responses between fresh and frozen samples from the same patients in **(a)** Karolinska ($n=10$) and **(b)** Oslo cohorts ($n=3$) across the 46 common drugs (the points with the same color). The correlation line (blue solid line) and the diagonal (grey dotted line) overlaps, indicating a good correlation, although there are some individual sample-specific differences between the fresh and frozen sample correlations.



Suppl. Fig. 14. Influence of blast percentage on drug sensitivity scores (DSS_2). Distribution of Spearman correlation coefficient of all 47 common drugs with blast percentage separately in each cohort with blast percentage available. The number of significant correlations was zero in each cohort.



Suppl. Fig. 15. (a) Left: PPCA was used to visualise the DSS_2 profiles of the patients from the four studies using a dataset where drugs with more than 20% missing DSS_2 values were removed leaving 35 drugs. After this the rest of the missing values were removed leaving 522 samples that are plotted. Right: Heatmap of DSS_2 responses based on the same dataset as for the PPCA across the four different labs (color bar at the top), where the columns correspond to the 522 patient samples and rows to 35 overlapping drugs. The low DSS_2 levels indicate low potency. (b) The same dataset of 35 drugs and 522 samples was used to perform batch effect correction using Combat. Left: PPCA was used to visualise the Combat scores of the patients from the four studies. The Combat correction effectively makes the cohorts overlap with each other. Right: Heatmap of the Combat scores across the four different labs (color bar at the top), where the columns correspond to the 522 patient samples and rows to 35 of the overlapping drugs. The high Combat levels indicate high drug sensitivity. After the Combat correction, the drug response clusters are not anymore driven by the laboratories.



Suppl. Fig. 16. Forest plot of the model coefficients and their 95% confidence intervals from the LMMs estimated based on 522 samples separately for each experimental condition, indicating how much of the variability in the Combat corrected DSS_2 levels can be explained by the specific experimental conditions; there is no significant difference in the Combat corrected DSS_2 values for any of the experimental variables. This suggests that variation arising from differences in experimental procedures has been effectively corrected; however, it is likely that genuine biological signals have been partially attenuated as well. ^aThe p-values are Bonferroni corrected and indicate the significance of the Combat corrected DSS_2 differences compared to the reference condition.