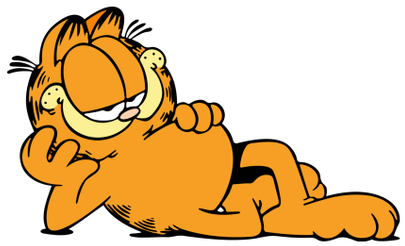
UGSRP: Blood Proportion Estimation on Epigenome Data

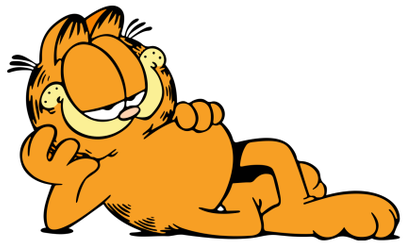
Olga Xu

Epigenetics: Above Genetics

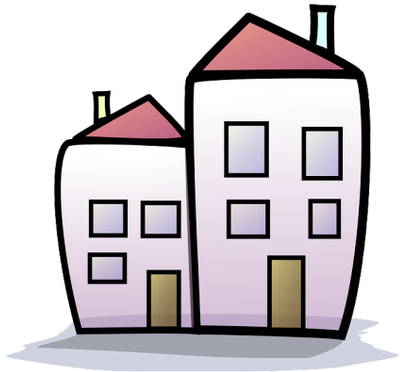
Epigenetics: Above Genetics



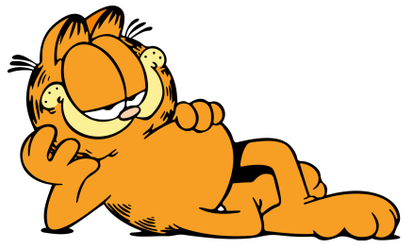
Epigenetics: Above Genetics



+



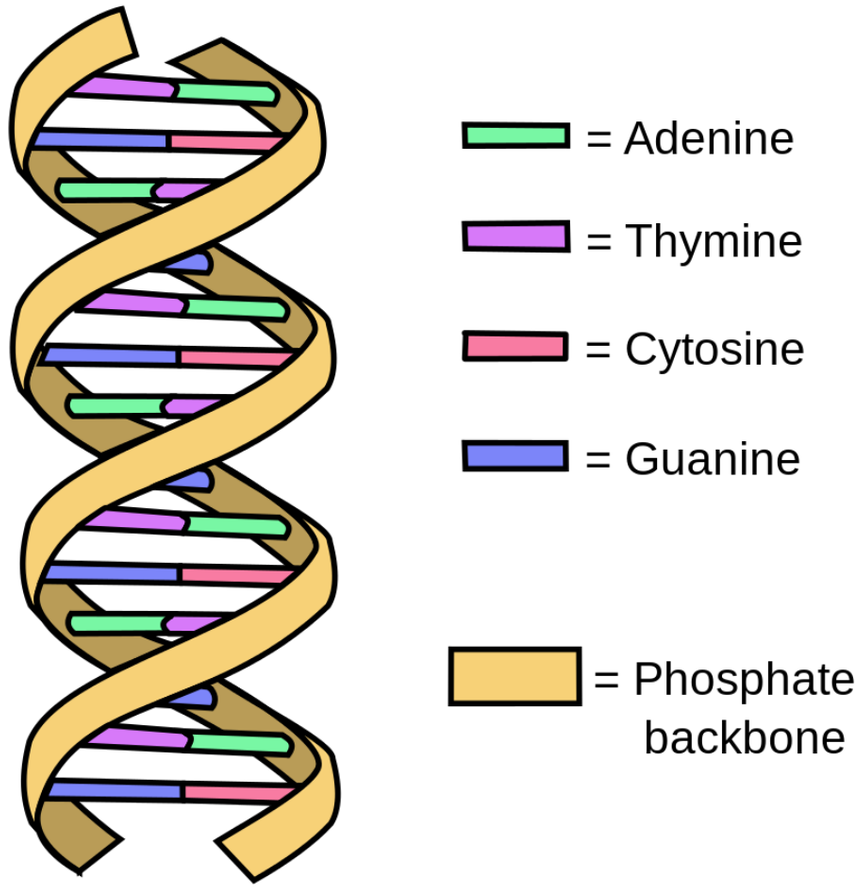
=



+

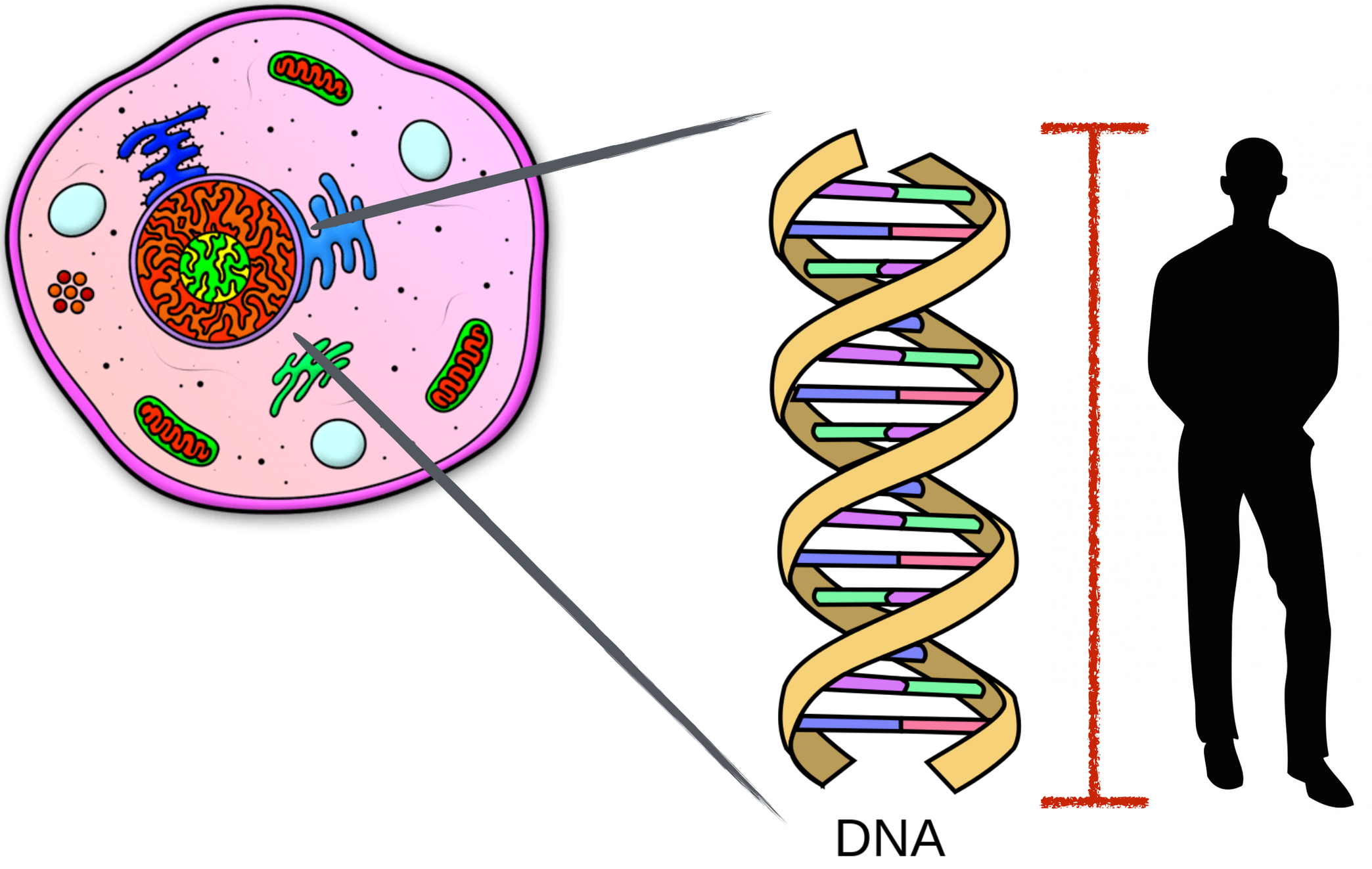
=

DNA Structure

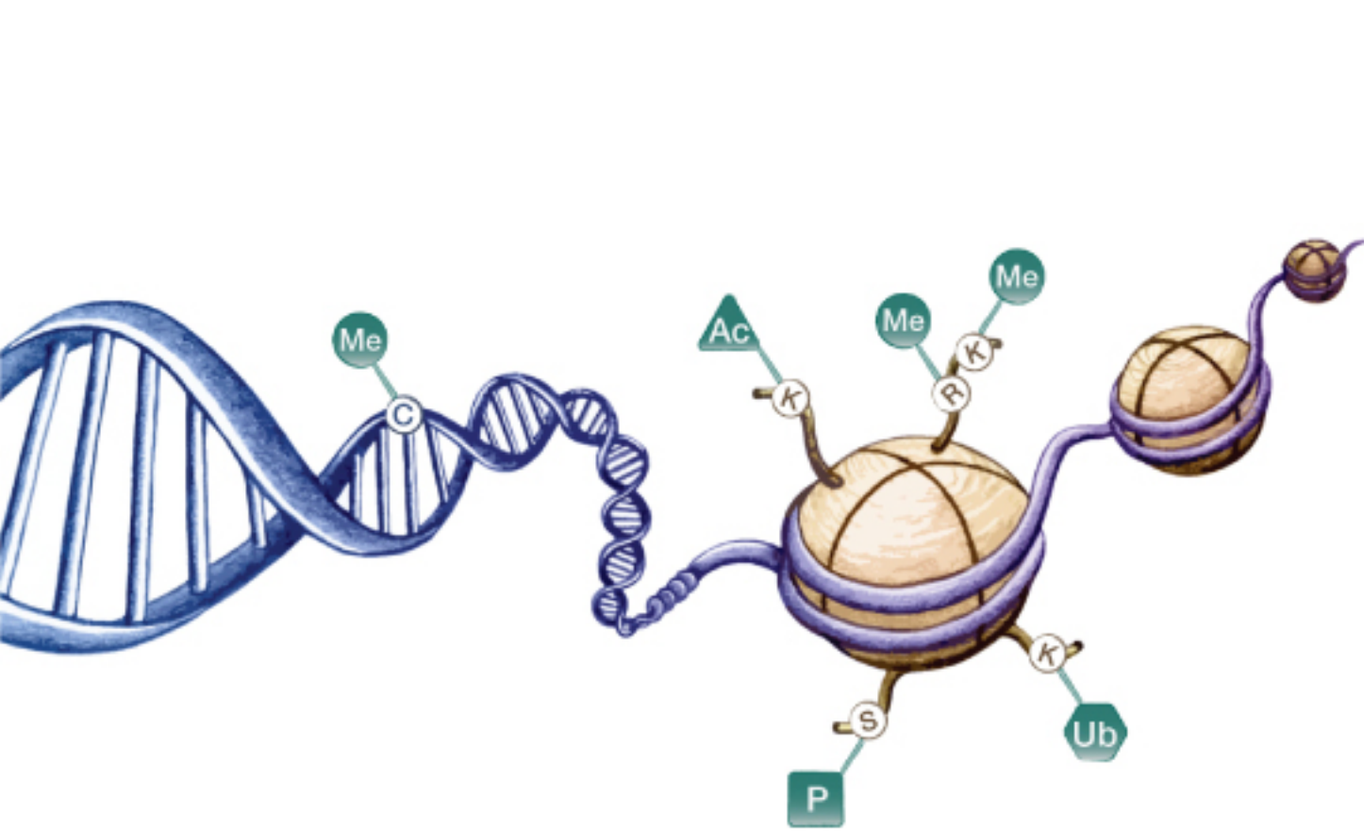


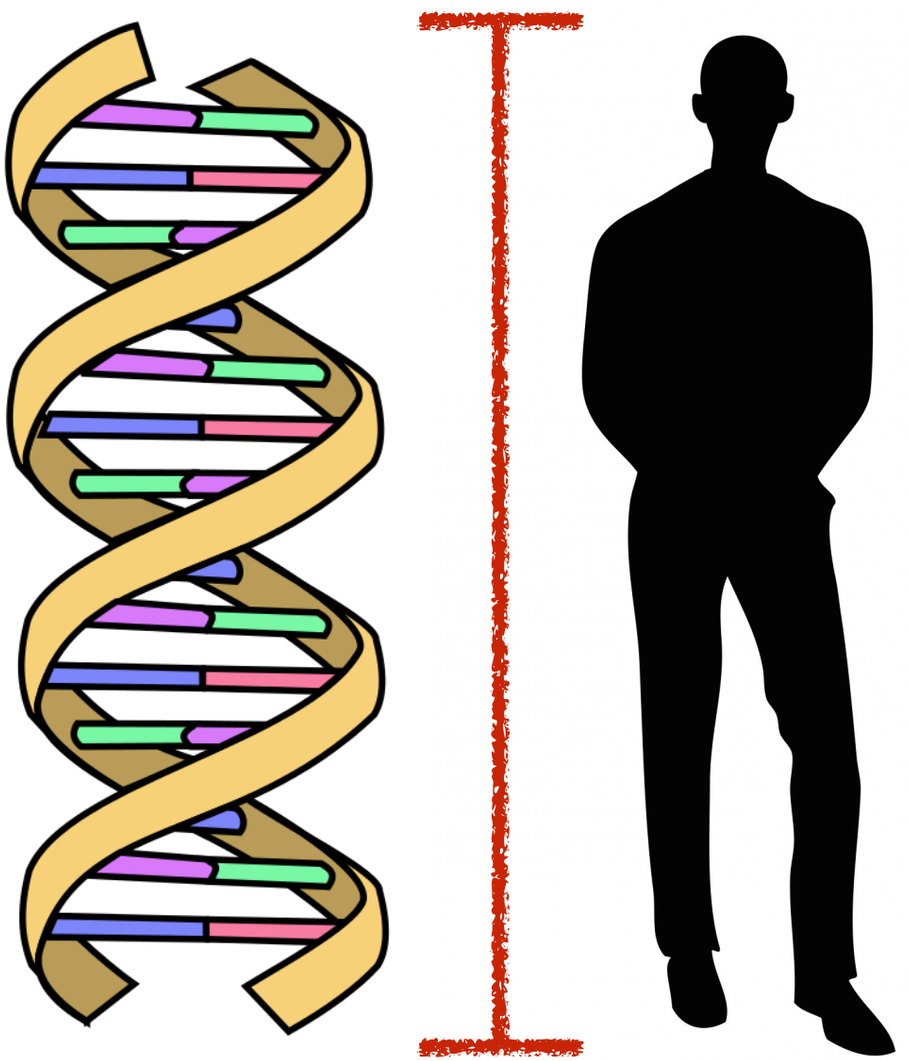
DNA

DNA Structure



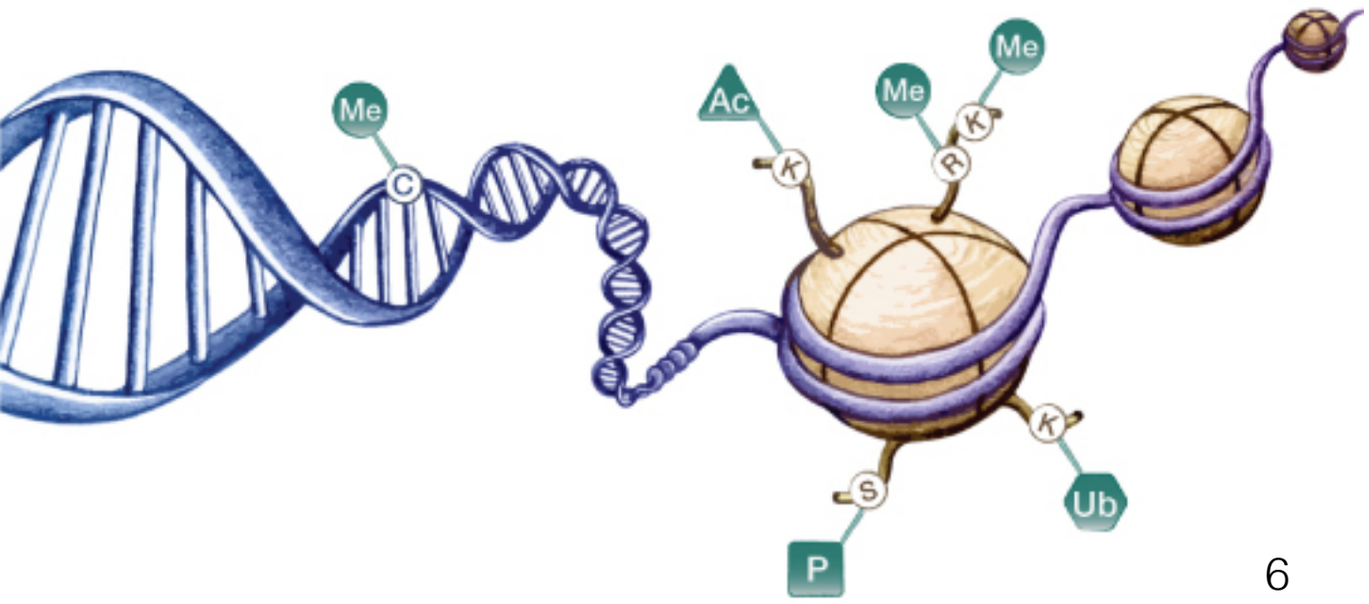
DNA Structure





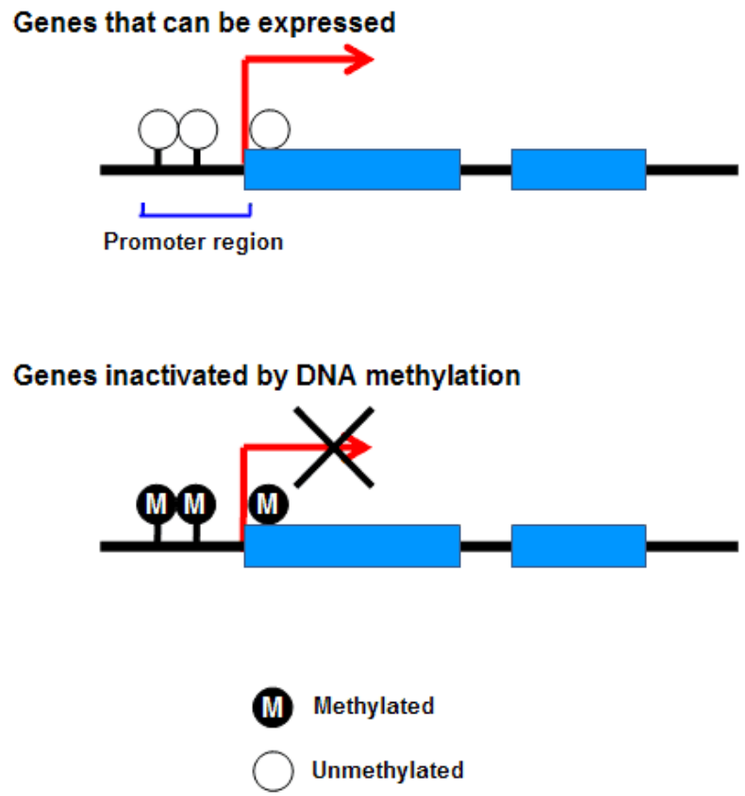
DNA

Epigenetics



DNA Methylation: Gene Silencing by Modification of DNA and Histones

DNA methylation: DNA structure can be modified by covalent attachment of methyl group.



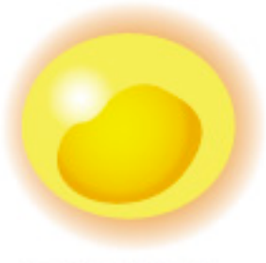
CpG sites: regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide in the linear sequence of bases along its length. DNA methylation often happens at the CpG sites.

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

Whole Blood DNA Methylation

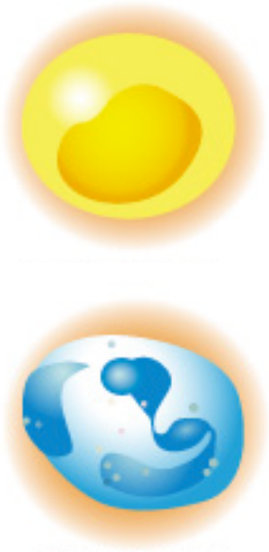
Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.



Lymphocytes: CD4T + CD8T + NK + Bcells

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.



Lymphocytes: CD4T + CD8T + NK + Bcells

Granulocytes

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

Lymphocytes: CD4T + CD8T + NK + Bcells

Granulocytes

Monocytes

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

CG_id	CD4+1	CD4+2	CD4+3	CD4+4	CD4+5	CD4+6
cg00000029	0.65	0.67	0.69	0.68	0.71	0.74
cg00000108	0.95	0.94	0.94	0.95	0.92	0.94
cg00000109	0.91	0.92	0.91	0.93	0.90	0.92

……

Table: 6 people's methylation profile on CD4Ts' cells

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

CG_id	CD4+1
cg00000029	0.65359

$$\beta = \frac{M}{M+U+\text{offset}}$$

CG_id	CD4+1	CD4+2	CD4+3	CD4+4	CD4+5	CD4+6
cg00000029	0.65	0.67	0.69	0.68	0.71	0.74
cg00000108	0.95	0.94	0.94	0.95	0.92	0.94
cg00000109	0.91	0.92	0.91	0.93	0.90	0.92

……

Table: 6 people's methylation profile on CD4Ts' cells

Whole Blood DNA Methylation

Whole Blood: heterogeneous collection of different cell types; each with a very different DNA methylation profile.

6 cell types and 6 males' sample

6 x 6 Samples

CG_id	CD4+1	CD4+2	CD4+3	CD4+4	CD4+5	CD4+6
cg00000029	0.65	0.67	0.69	0.68	0.71	0.74
cg00000108	0.95	0.94	0.94	0.95	0.92	0.94
cg00000109	0.91	0.92	0.91	0.93	0.90	0.92

.....

450000 CpG Probes

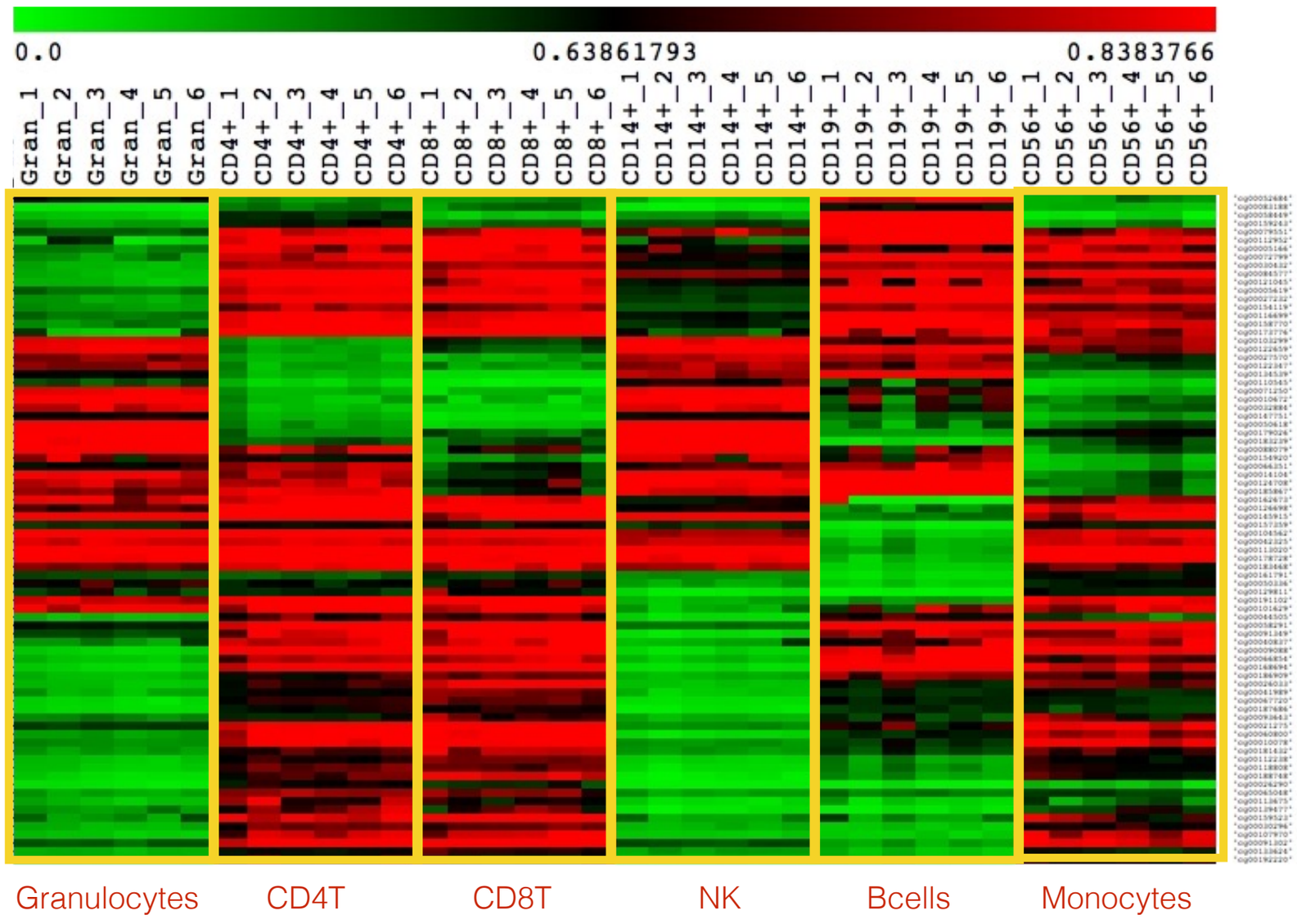
CG_id	CD4+1
cg00000029	0.65359

$$\beta = \frac{M}{M+U+\text{offset}}$$

Table: 6 people's methylation profile on CD4Ts' cells

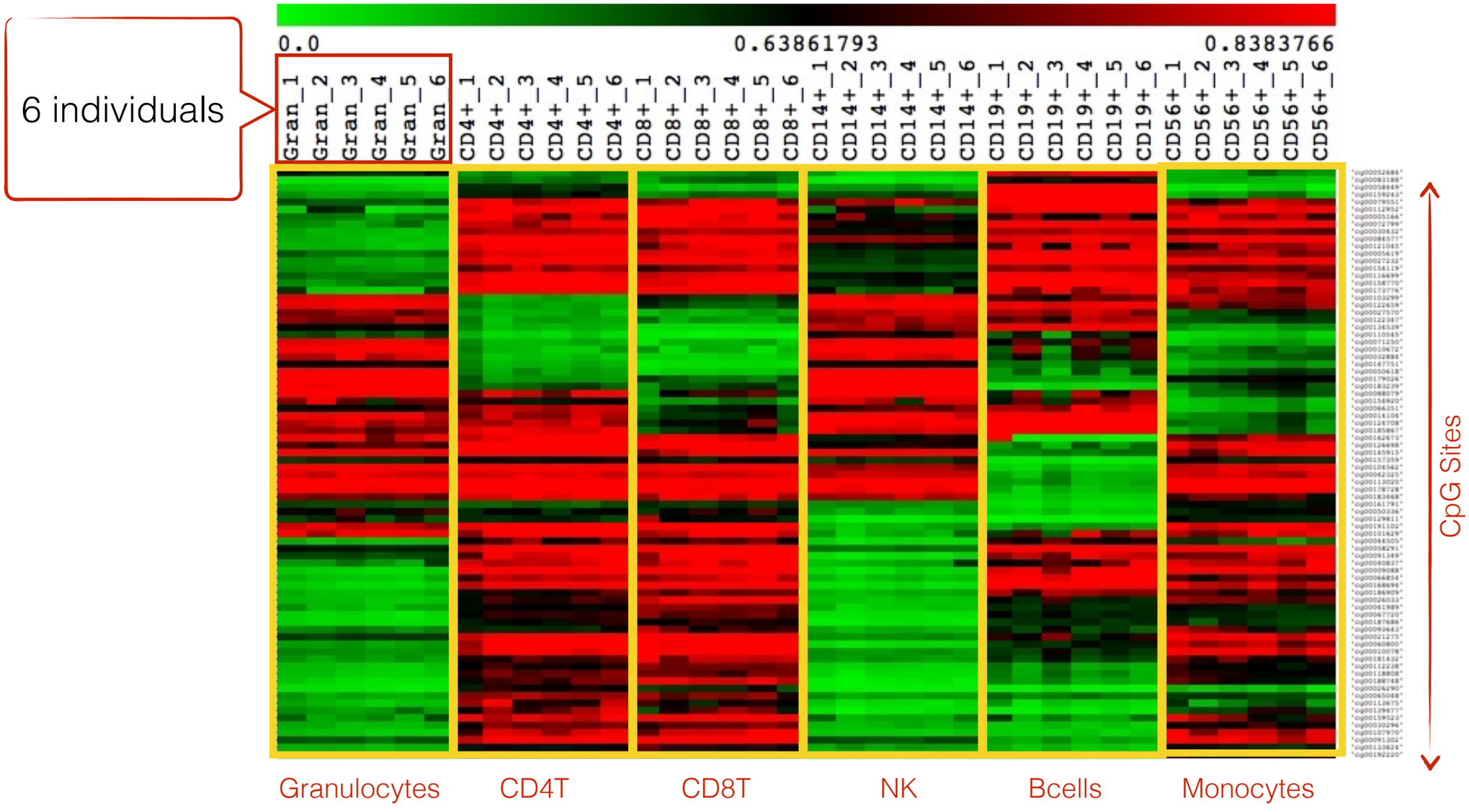
Whole Blood DNA Methylation

A heatmap of the cell sorted data shows very clear and consistent DNAm profiles for each cell type.



Whole Blood DNA Methylation

A heatmap of the cell sorted data shows very clear and consistent DNAm profiles for each cell type.



Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous

Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous

Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous

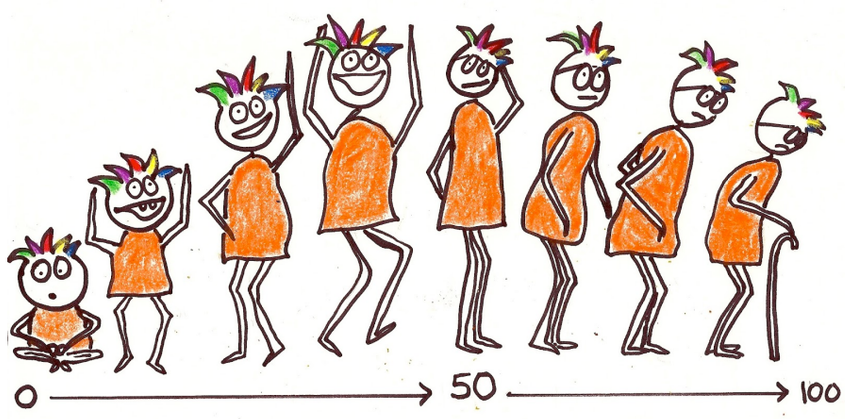
> Accounting for cellular heterogeneity is critical in epigenome-wide association studies
>
> Jaffe & Irrizary

Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous

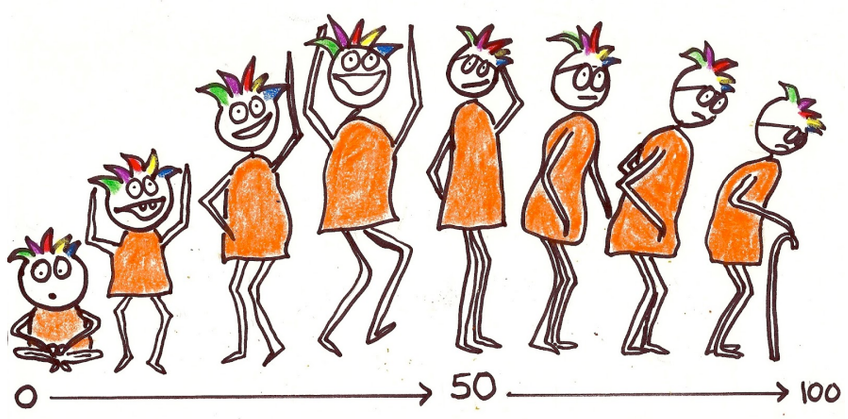
Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous



Project Goal

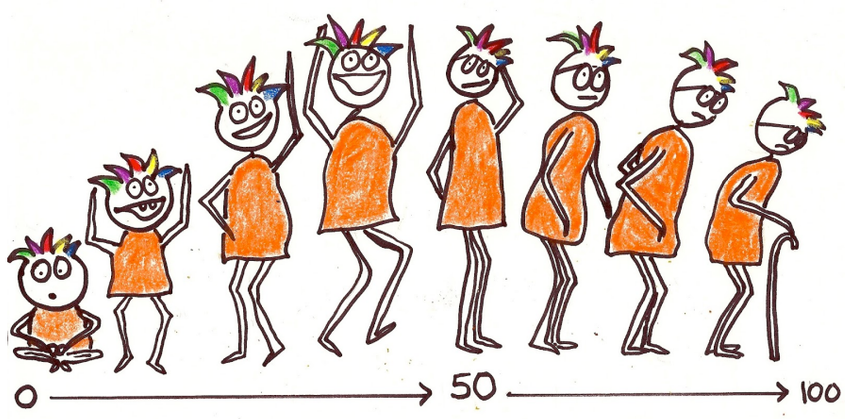
- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous



age

Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous



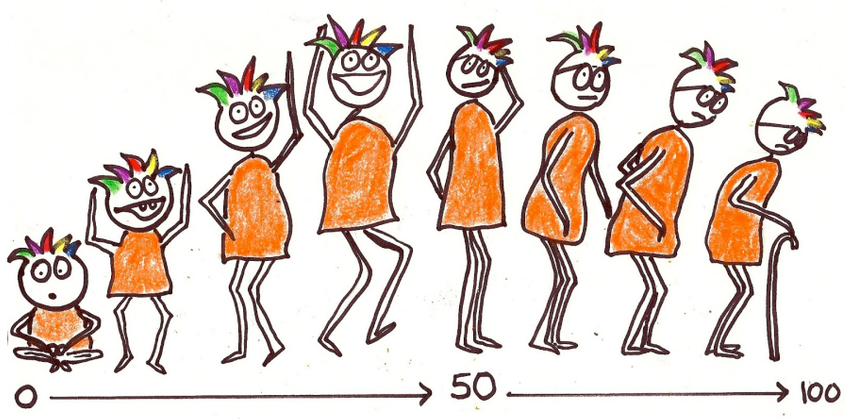
age

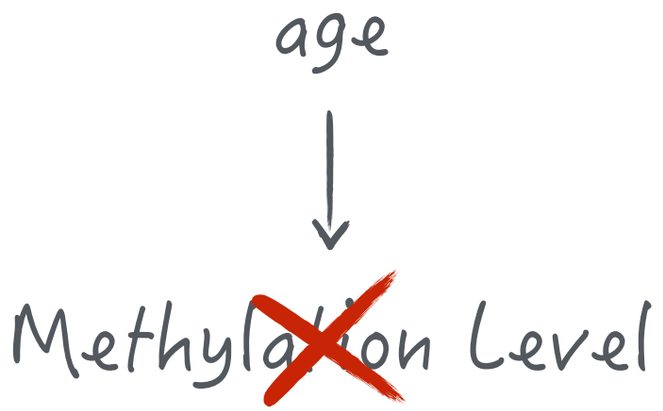
↓

Methylation Level

Project Goal

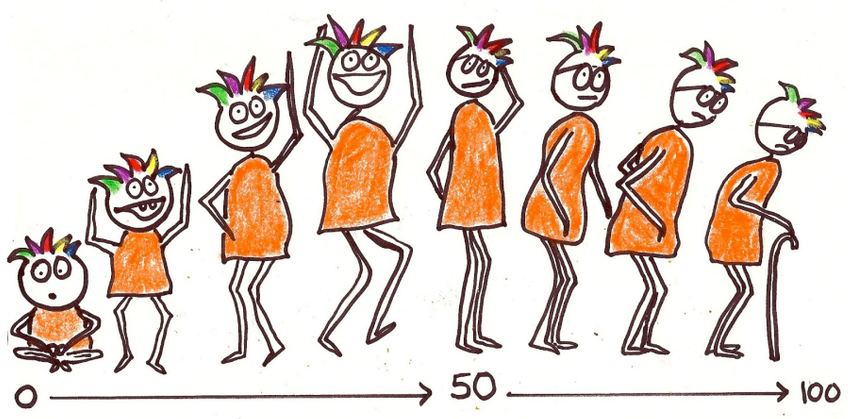
- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous





Project Goal

- **To estimate the whole blood proportion.**
- Want to diagnose diseases.
- When we are looking for differences in methylation due to disease, we need to make sure that what we are finding are not merely differences in cell proportion.
- Such failure to account for blood composition is very dangerous



age

↓

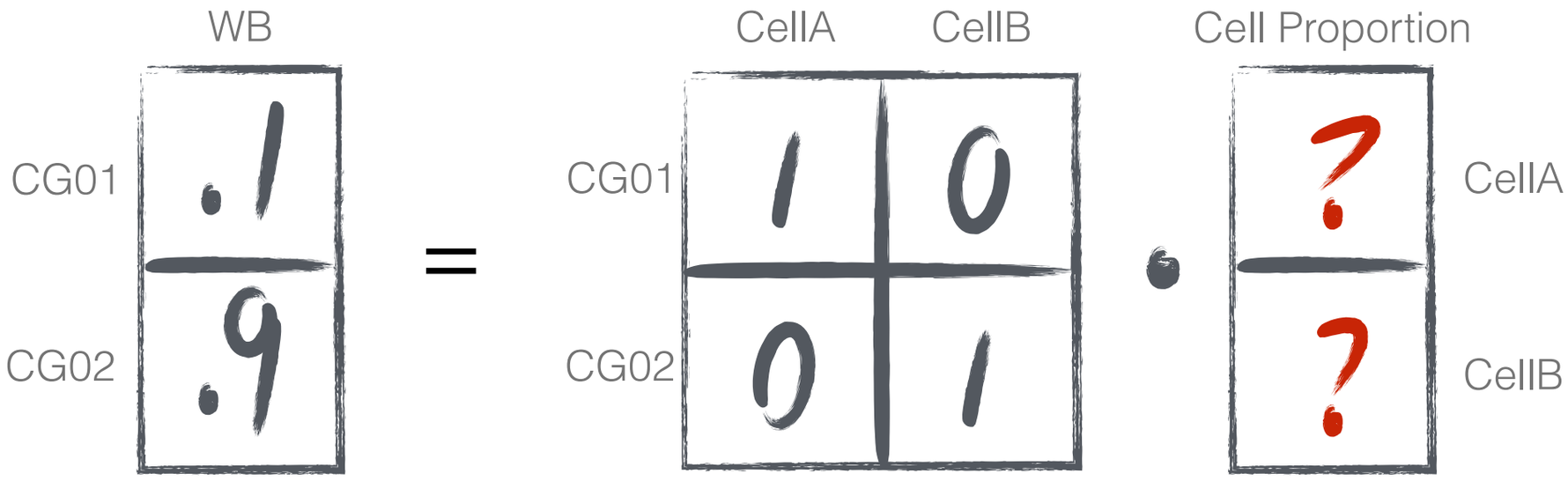
Whole Blood Composition

Estimate Whole Blood Proportion

Approach:

Assuming everyone has the same methylation profile, then a group of six non-parallel CpG site methylation profiles would provide me the blood proportion estimate.
(simple linear model: Ax = b)

People have different methylation profile. However, for every six non-parallel CpG methylation profiles would give a reasonable estimate of the whole blood composition.

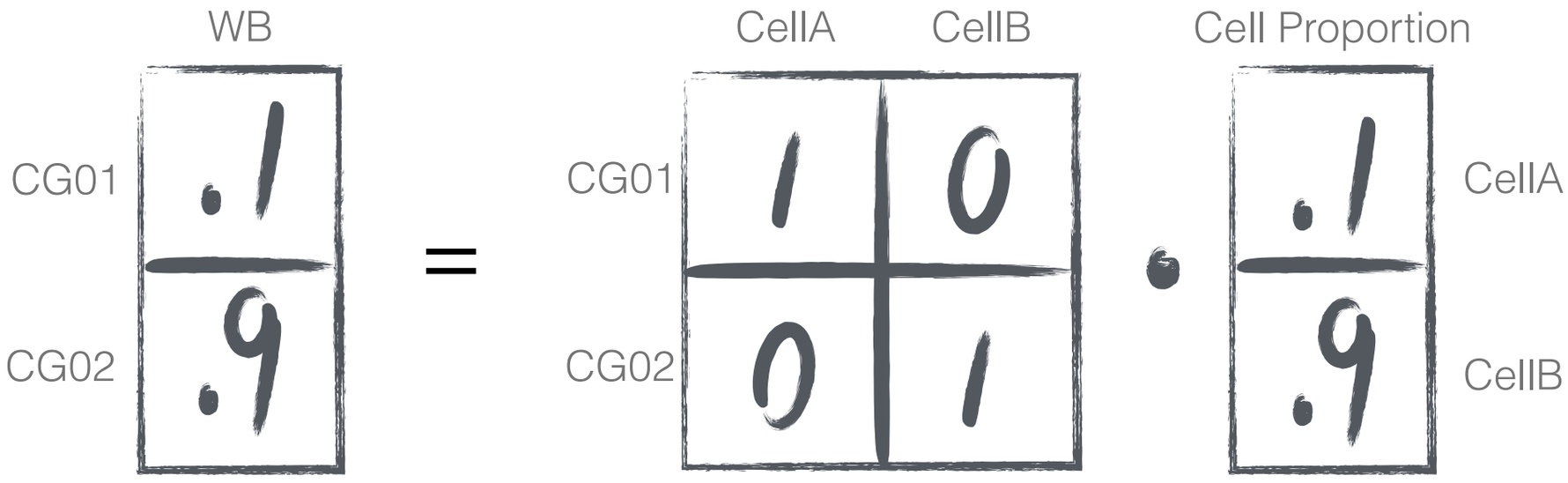


Estimate Whole Blood Proportion

Approach:

Assuming everyone has the same methylation profile, then a group of six non-parallel CpG site methylation profiles would provide me the blood proportion estimate.
(simple linear model: Ax = b)

People have different methylation profile. However, for every six non-parallel CpG methylation profiles would give a reasonable estimate of the whole blood composition.



Single Matrix Example

	Gran	CD4T	CD8T	NK	Bcell	Mono
cg00005166	0.32	0.79	0.84	0.64	0.68	0.76
cg00005619	0.35	0.82	0.82	0.51	0.87	0.76
cg00009088	0.11	0.90	0.94	0.09	0.83	0.89
cg00010078	0.28	0.85	0.87	0.26	0.52	0.85
cg00010672	0.94	0.20	0.17	0.91	0.54	0.35
cg00014104	0.84	0.77	0.51	0.87	0.88	0.36

Single Matrix Example

avg (

	Gran_1	Gran_2	Gran_3	Gran_4	Gran_5	Gran_6
cg00005166	0.3446111	0.345826	0.3227671	0.2207103	0.3226565	0.3514021

)

	Gran	CD4T	CD8T	NK	Bcell	Mono
cg00005166	0.32	0.79	0.84	0.64	0.68	0.76
cg00005619	0.35	0.82	0.82	0.51	0.87	0.76
cg00009088	0.11	0.90	0.94	0.09	0.83	0.89
cg00010078	0.28	0.85	0.87	0.26	0.52	0.85
cg00010672	0.94	0.20	0.17	0.91	0.54	0.35
cg00014104	0.84	0.77	0.51	0.87	0.88	0.36

Single Matrix Example

avg (

	Gran_1	Gran_2	Gran_3	Gran_4	Gran_5	Gran_6
cg00005166	0.3446111	0.345826	0.3227671	0.2207103	0.3226565	0.3514021

)

	WB
cg00005166	0.53
cg00005619	0.55
cg00009088	0.42
cg00010078	0.48
cg00010672	0.69
cg00014104	0.76

	Gran	CD4T	CD8T	NK	Bcell	Mono
cg00005166	0.32	0.79	0.84	0.64	0.68	0.76
cg00005619	0.35	0.82	0.82	0.51	0.87	0.76
cg00009088	0.11	0.90	0.94	0.09	0.83	0.89
cg00010078	0.28	0.85	0.87	0.26	0.52	0.85
cg00010672	0.94	0.20	0.17	0.91	0.54	0.35
cg00014104	0.84	0.77	0.51	0.87	0.88	0.36

Single Matrix Example

avg (

	Gran_1	Gran_2	Gran_3	Gran_4	Gran_5	Gran_6
cg00005166	0.3446111	0.345826	0.3227671	0.2207103	0.3226565	0.3514021

)

	WB
cg00005166	0.53
cg00005619	0.55
cg00009088	0.42
cg00010078	0.48
cg00010672	0.69
cg00014104	0.76

=

	Gran	CD4T	CD8T	NK	Bcell	Mono
cg00005166	0.32	0.79	0.84	0.64	0.68	0.76
cg00005619	0.35	0.82	0.82	0.51	0.87	0.76
cg00009088	0.11	0.90	0.94	0.09	0.83	0.89
cg00010078	0.28	0.85	0.87	0.26	0.52	0.85
cg00010672	0.94	0.20	0.17	0.91	0.54	0.35
cg00014104	0.84	0.77	0.51	0.87	0.88	0.36

	Cell Prop.
Gran	0.5
CD4T	0.1
CD8T	0.1
NK	0.1
Bcell	0.1
Mono	0.1

Project Goal: Estimate Whole Blood Proportion

Preprocesses:

Probes Selection: hypothesis testing ($P < 0.05$), methylation range test, individual variance test

Matrices Selection: singularity testing (condition number), accuracy testing

Project Goal: Estimate Whole Blood Proportion

Preprocesses:

Probes Selection: hypothesis testing ($P < 0.05$), methylation range test, individual variance test

Matrices Selection: singularity testing (condition number), accuracy testing

1	0
0	1

1	0
0.9999	0.0001

Condition number for the left matrix is 1. Condition number for the right matrix is 19998.

Project Goal: Estimate Whole Blood Proportion

Preprocesses:

Probes Selection: hypothesis testing ($P < 0.05$), methylation range test, individual variance test

Matrices Selection: singularity testing (condition number), accuracy testing

1	0
0	1

1	0
0.9999	0.0001

Condition number for the left matrix is 1. Condition number for the right matrix is 19998.

Parameter Optimization:

Grid Search: optimize the parameters of a modelling with an internal cross-validation

Project Goal: Estimate Whole Blood Proportion

Preprocesses:

Probes Selection: hypothesis testing ($P < 0.05$), methylation range test, individual variance test

Matrices Selection: singularity testing (condition number), accuracy testing

1	0
0	1

1	0
0.9999	0.0001

Condition number for the left matrix is 1. Condition number for the right matrix is 19998.

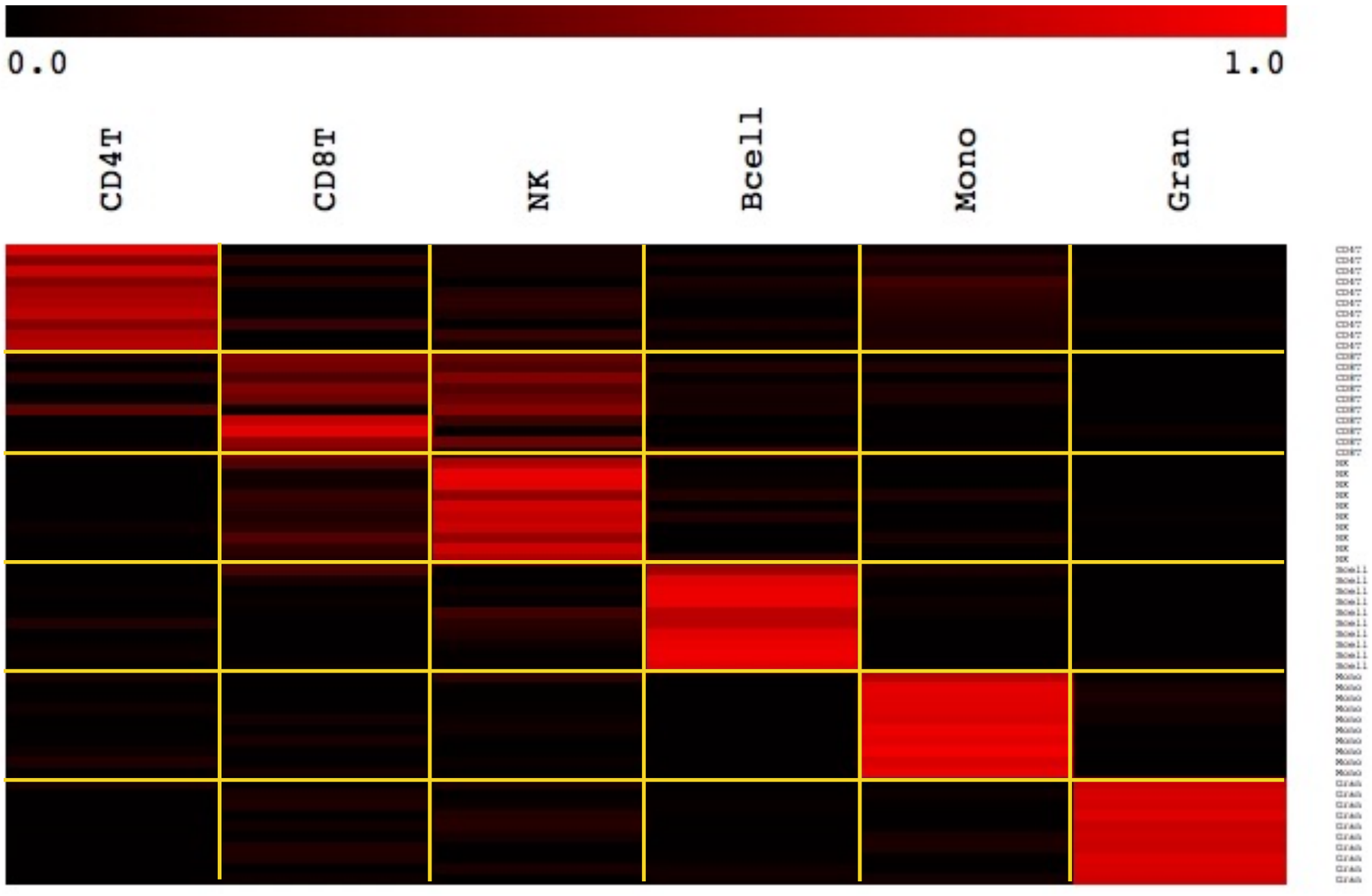
Parameter Optimization:

Grid Search: optimize the parameters of a modelling with an internal cross-validation

Validation:

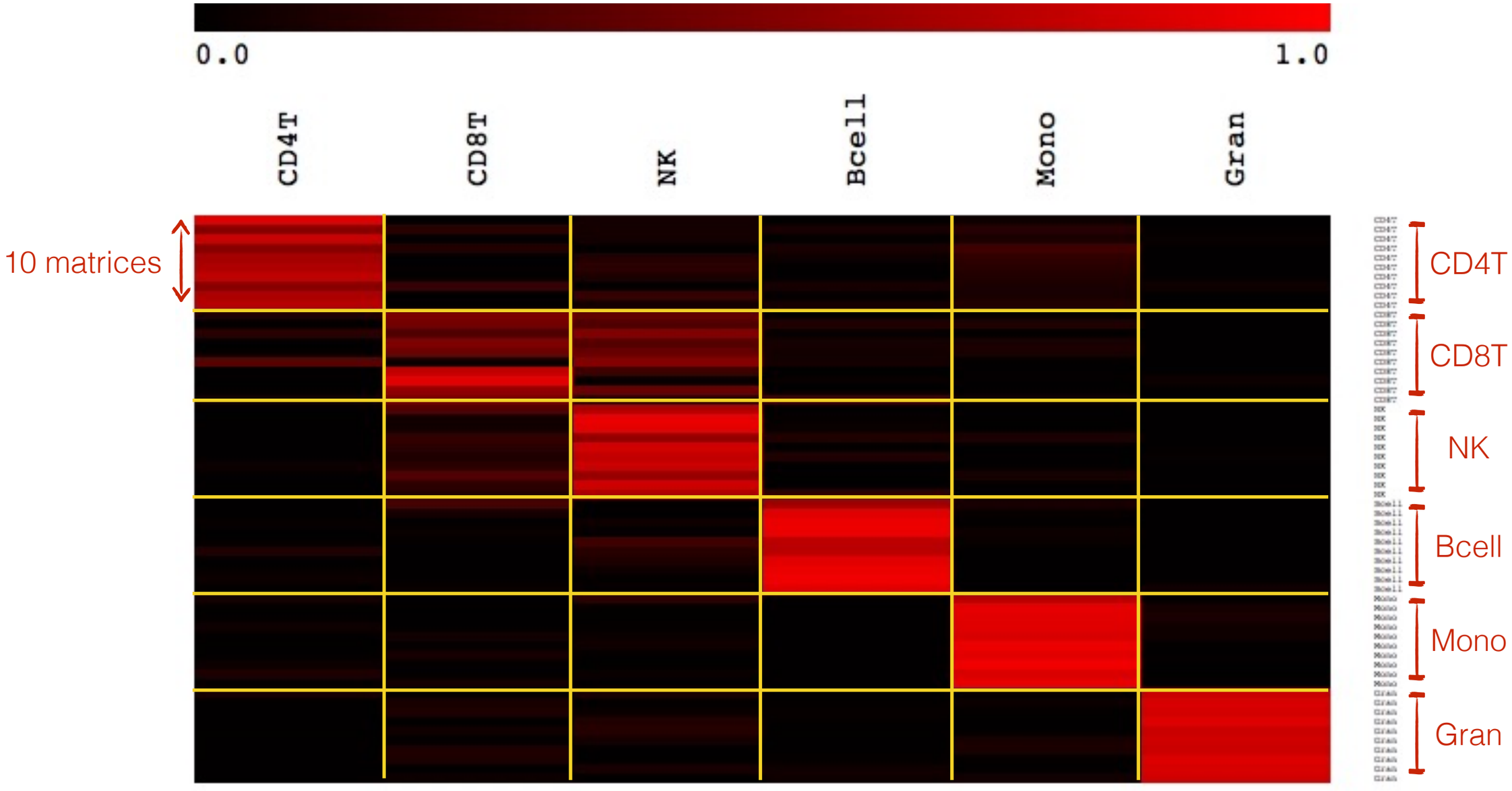
Cross Validation: 6-fold leave-one validation

Single Matrix Prediction



10 matrices individual estimation accuracies.

Single Matrix Prediction



10 matrices individual estimation accuracies.

Leave-One Validation Results with All Preprocesses

	CD4T	CD8T	NK	Bcell	Mono	Gran
CD4T	80%	2%	5%	0%	11%	2%
CD8T	2%	70%	28%	0%	0%	1%
NK	0%	9%	91%	0%	0%	0%
Bcell	0%	0%	3%	97%	0%	0%
Mono	2%	0%	0%	0%	97%	1%
Gran	0%	5%	1%	1%	0%	93%

Leave-One Validation Results with All Preprocesses

	CD4T	CD8T	NK	Bcell	Mono	Gran
CD4T	80%	2%	5%	0%	11%	2%
CD8T	2%	70%	28%	0%	0%	1%
NK	0%	9%	91%	0%	0%	0%
Bcell	0%	0%	3%	97%	0%	0%
Mono	2%	0%	0%	0%	97%	1%
Gran	0%	5%	1%	1%	0%	93%

Confusion matrix for estimating each cell with 100 matrices combined. Average accuracy combining 100 matrices: 88%.

Iterative Learning: The Greedy Approach

Problem:

Some cell types' estimation is significantly better than the others.

Iterative Learning:

Solve cell proportions one at each step.

Fix the best proportion estimation result for a cell type and make smaller matrix to estimate the remaining cell proportions. Then fix the best estimation and estimate for the rest.

Iterative Learning: The Greedy Approach

Problem:

Some cell types' estimation is significantly better than the others.

Iterative Learning:

Solve cell proportions one at each step.

Fix the best proportion estimation result for a cell type and make smaller matrix to estimate the remaining cell proportions. Then fix the best estimation and estimate for the rest.

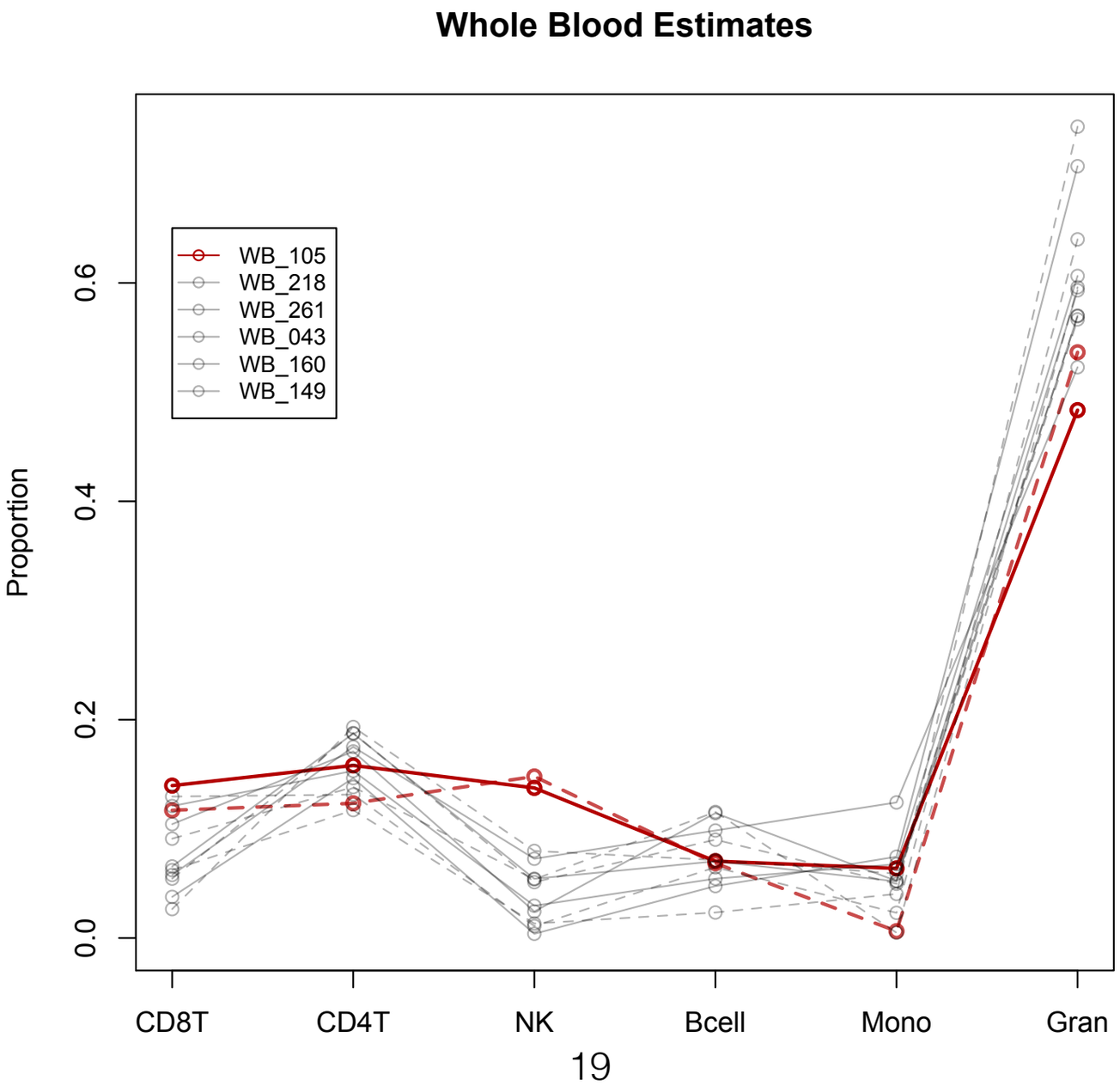
	CD4T	CD8T	NK	Bcell	Mono	Gran
CD4T	90%	1%	3%	0%	4%	2%
CD8T	3%	87%	9%	0%	0%	1%
NK	2%	6%	92%	0%	0%	0%
Bcell	0%	0%	3%	97%	0%	0%
Mono	1%	0%	0%	0%	98%	1%
Gran	0%	3%	1%	0%	0%	96%

Average accuracy: 93.3%.

Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

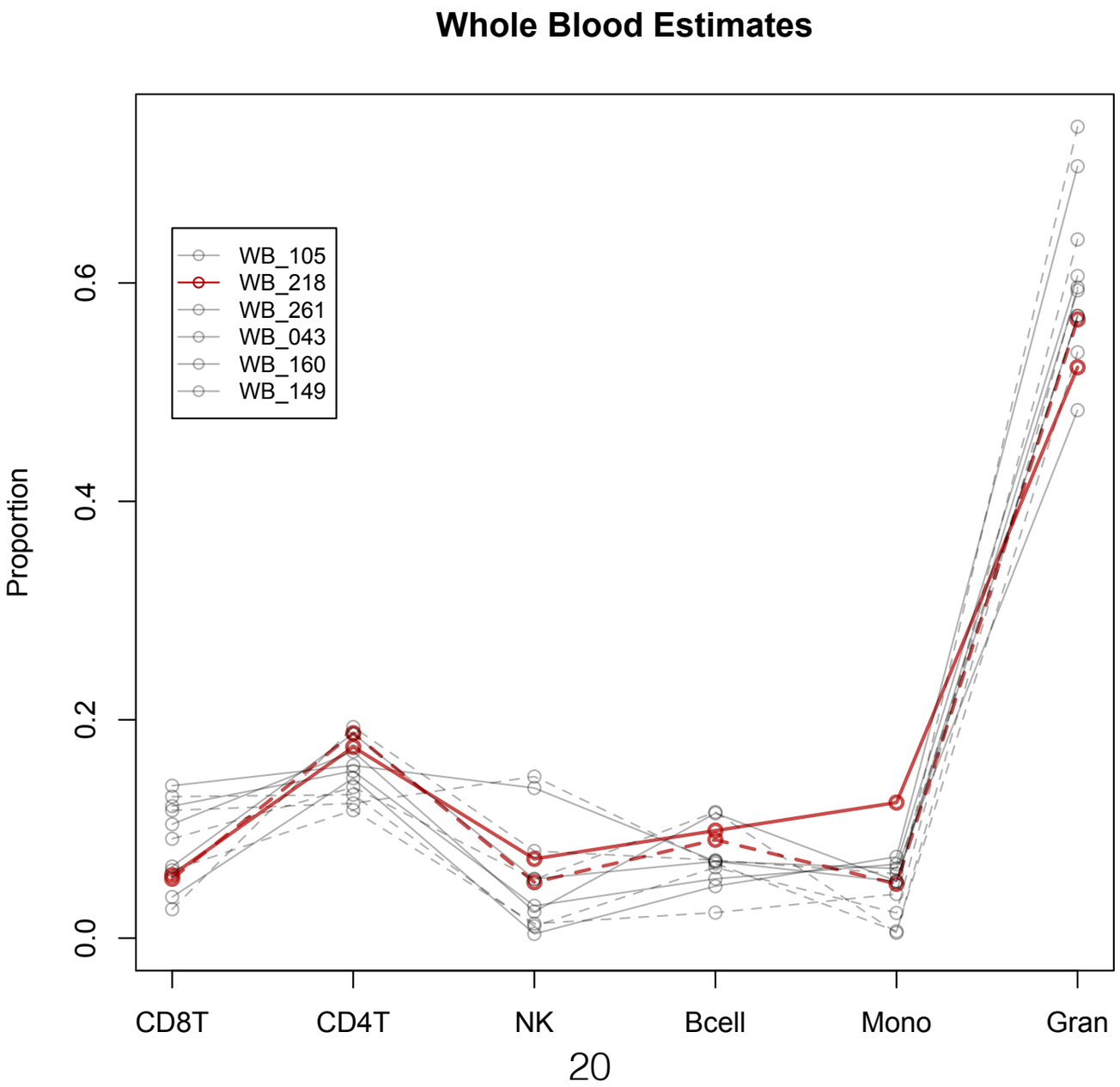
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

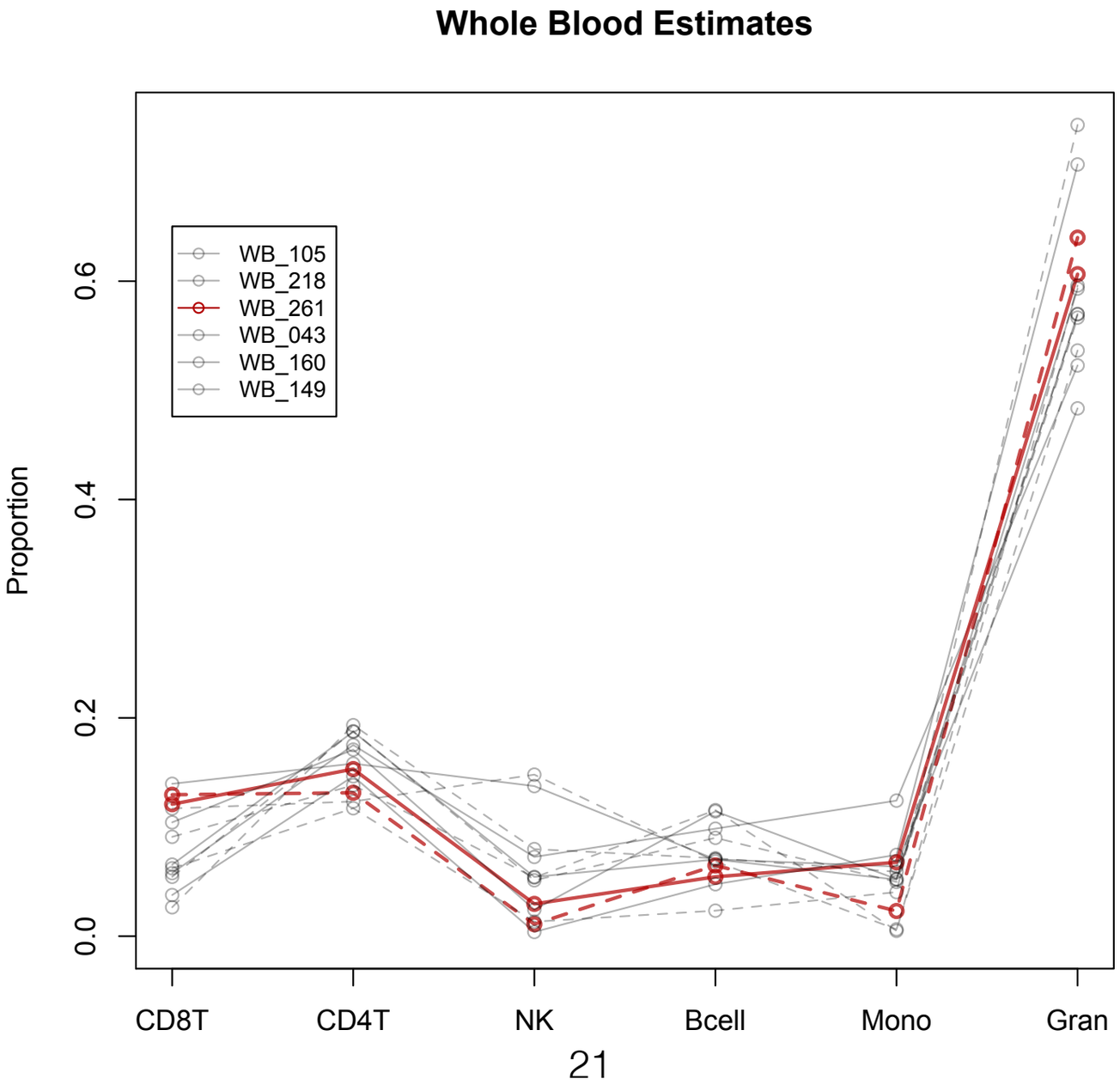
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

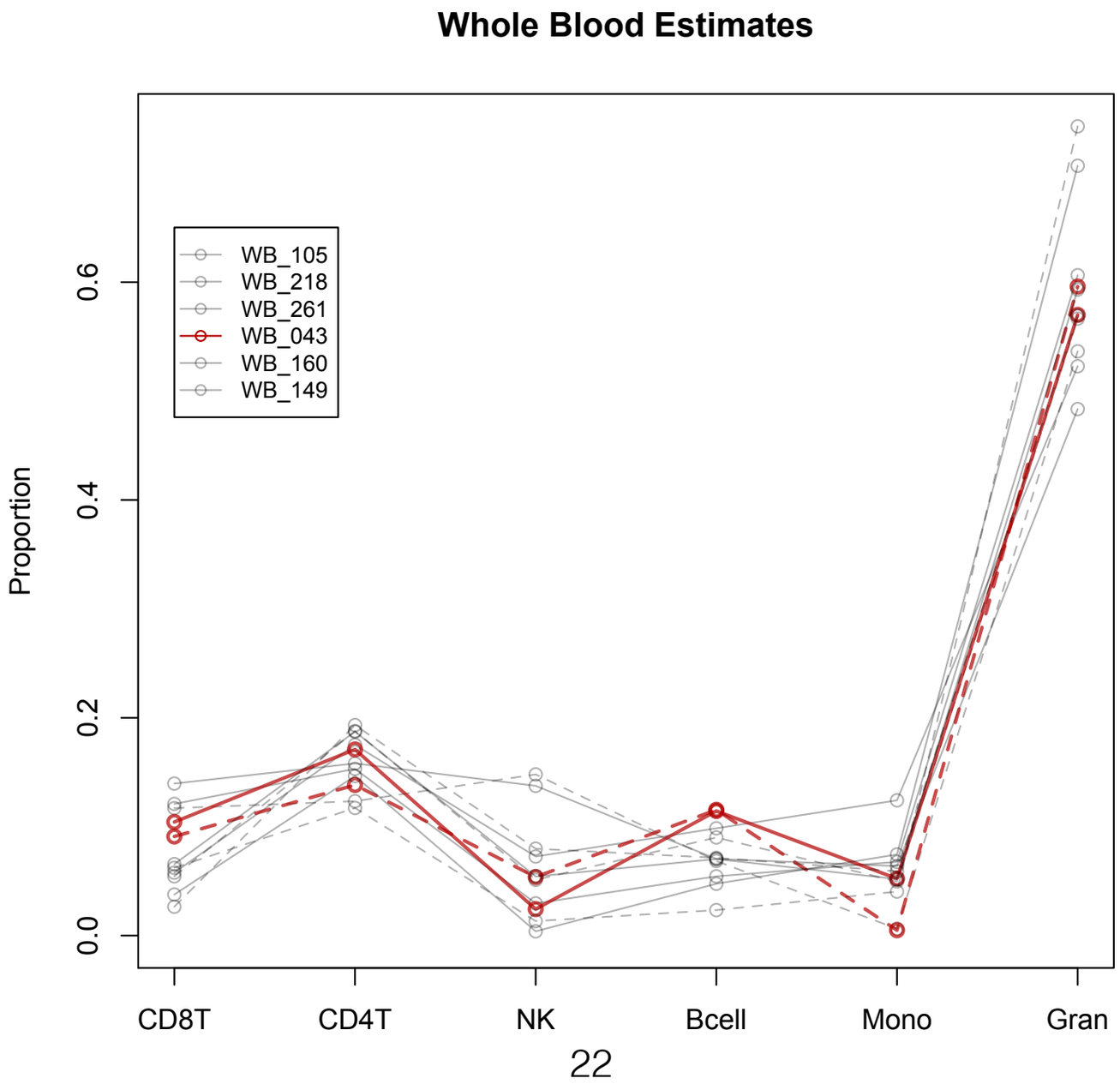
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

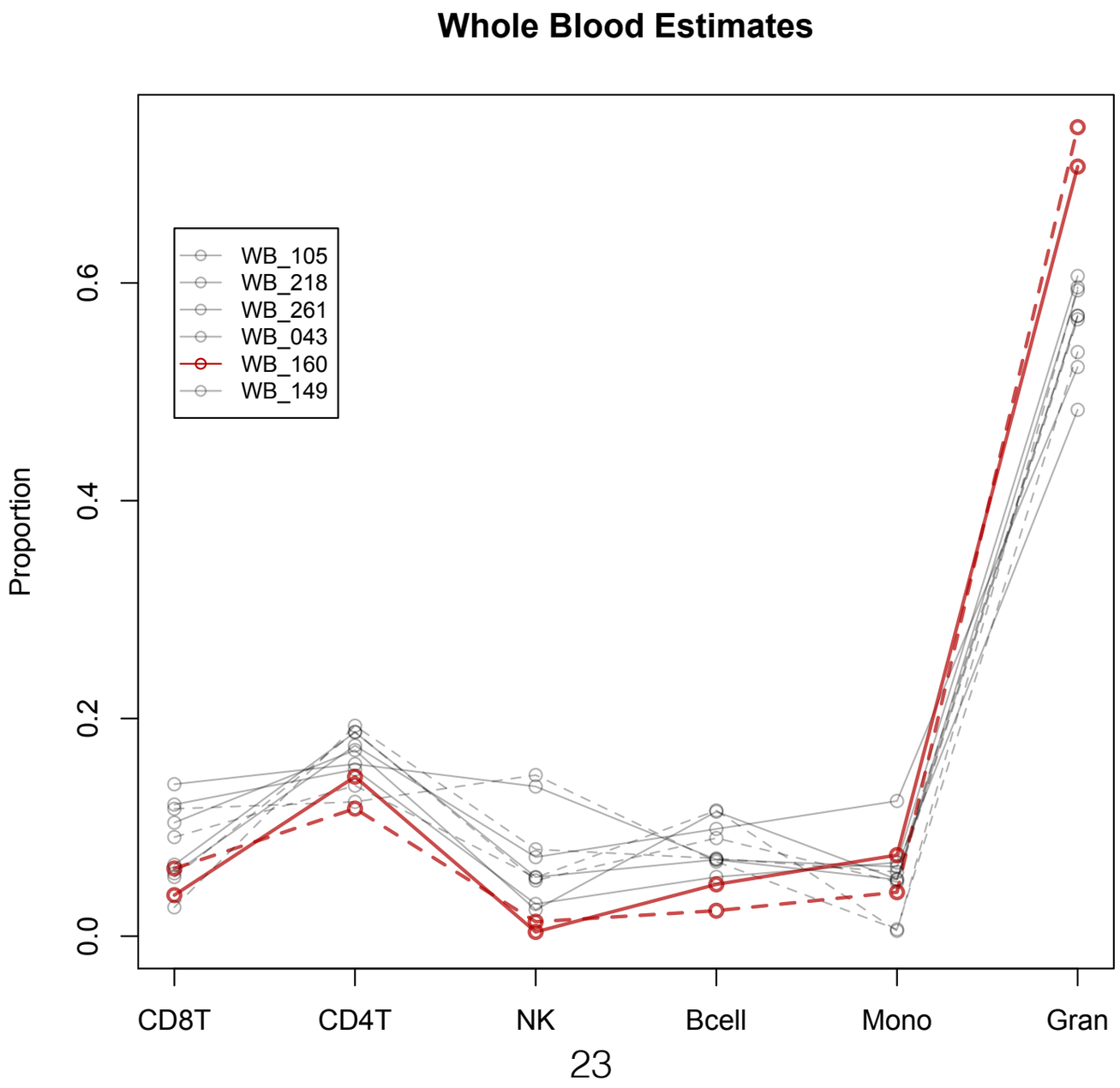
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

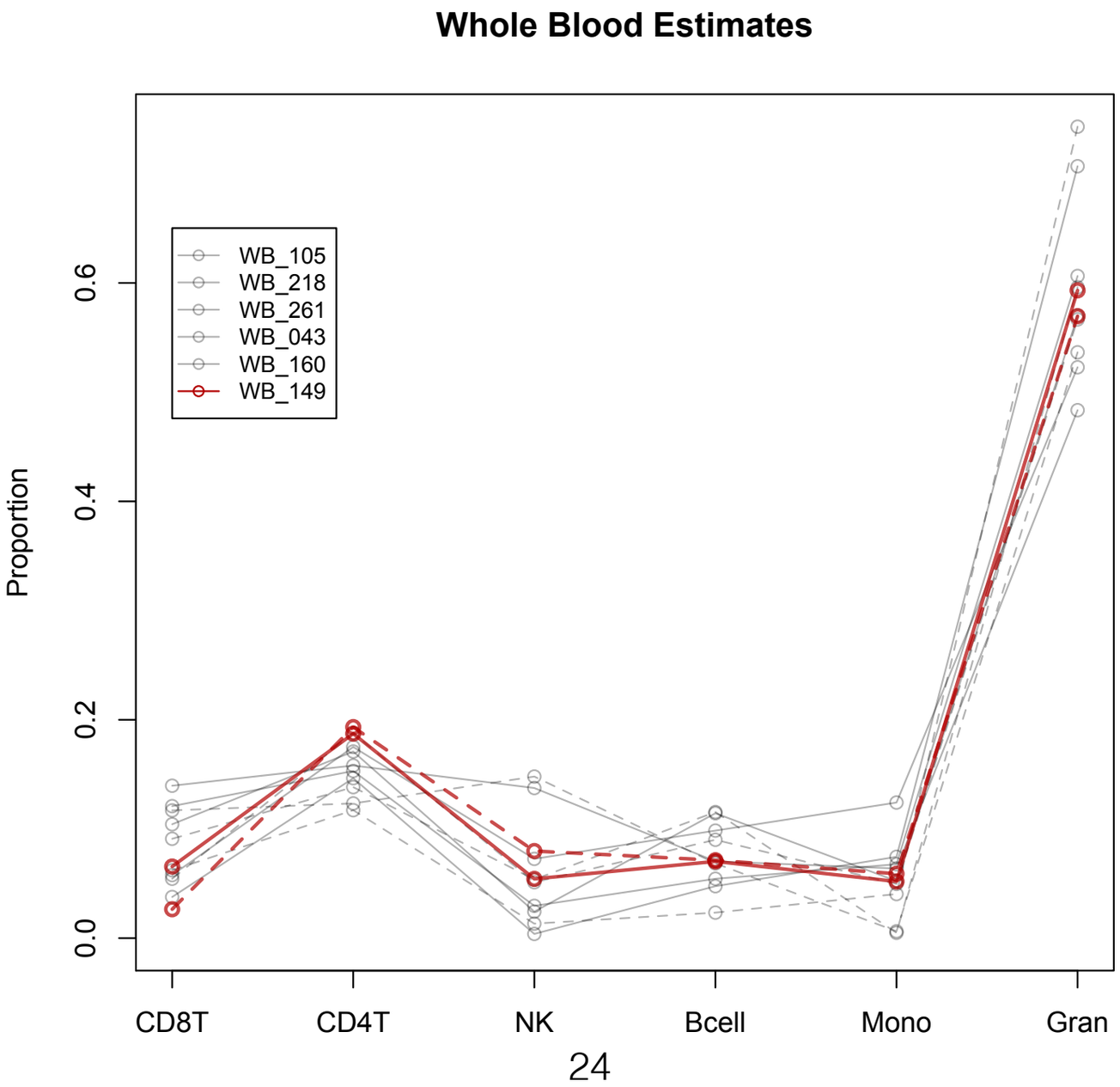
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

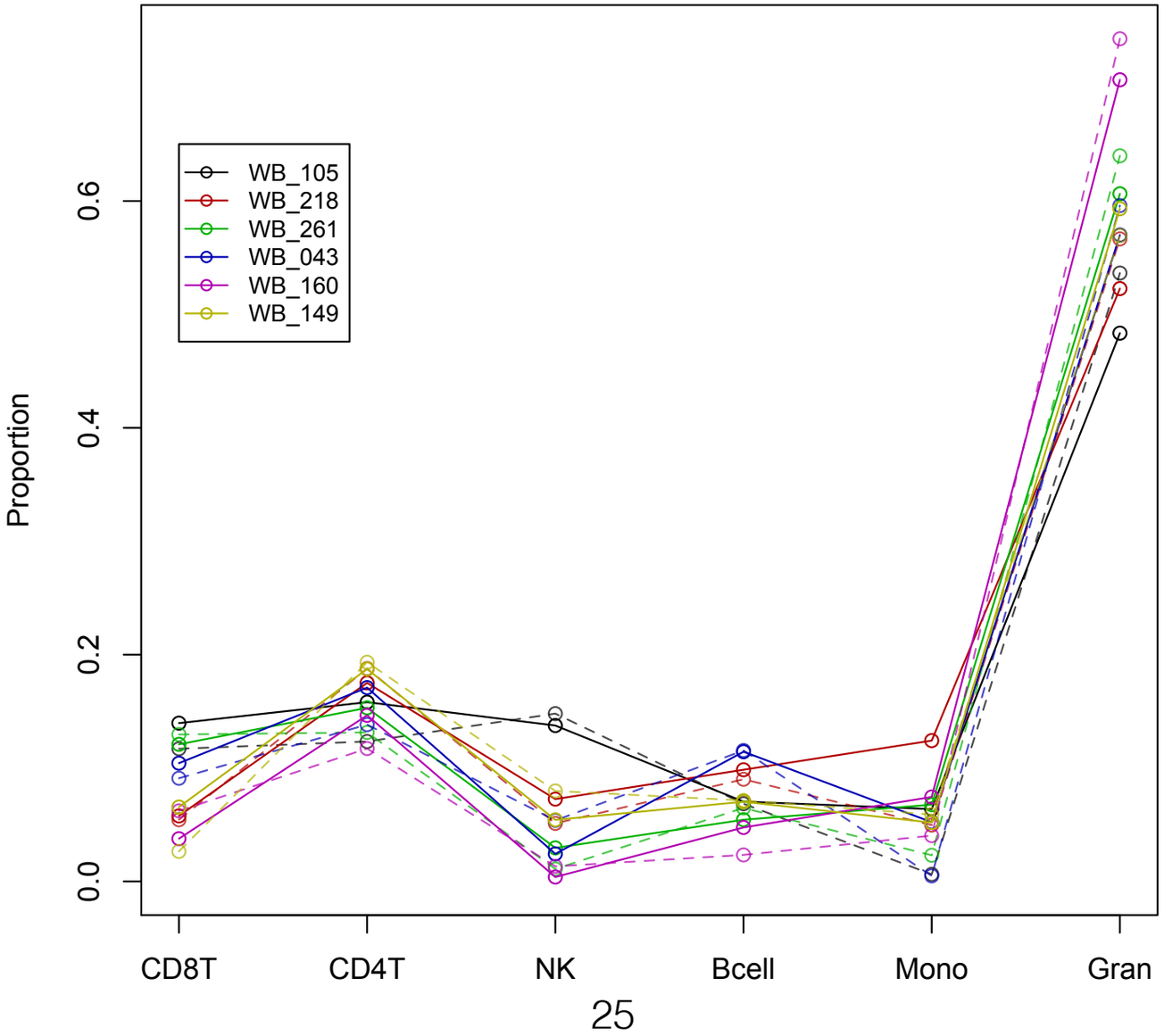
Below are comparisons of my model and the minfi package's model on six whole blood samples.



Comparing to the Minfi Package's Results

Minfi package was developed to tackle the same problem using linear regression model.

Below are comparisons of my model and the minfi package's model on six whole blood samples.



Minfi vs. Current Method

Minfi

Easier and less computational costly to build the model.

Quick in analysis.

Requires 560+ CpG methylation data.

Can adapt only whole blood data.

Accepts only raw data.

Current Method

Harder and more computational costly to build the model.

Quick in analysis.

Requires 100+ CpG methylation data.

Can adapt any form of heterogeneous cell data.

Accepts both raw and processed data.

Conclusion

Importance of considering cell composition variability in epigenetic studies.

Final Accuracy: 93%

Acknowledgement

Supervisor

M. Brudno

Mentor

A. Turinsky

Research Organization

Undergrad Summer Research Group (UGSRP)

Whole Blood Data Provider

Sickkid's Weksberg Lab

Special Thanks To Those Who Helped Me Preparing for This Presentation:

Michael Brudno's Group
A. Turinsky
P. Shao
X. Zeng
J. Chung
S. Betancourt
P. Luo