Bizonytalan Genotipizálás

Felkészülés

A mérés során bizonytalan genotipizálási adatok osztályozását fogjuk vizsgálni. A mérésre fel kell készülni. A szükséges ismeretek megtalálhatók azt útmutató második felében, valamint a Neurális Hálózatok (Altrichter-Horváth-Pataki-Strausz-Valyon) c. könyv 94-110-adik és 401-407-edik oldalán (MLP és PCA). Valamint segít a felkészülésben, ha megismerkednek a Matlab használatával, szintaxisával. Mindenképpen nyissa meg az adat fájlt Excel-ben, hogy láthassa felépítését, valamint tartalmának jellegét!

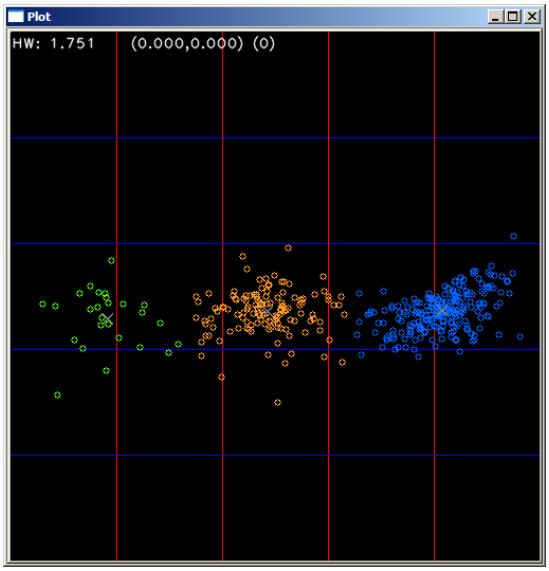
A mérésre házi feladattal kell készülni, és a mérés elején a kiadott melléklet alapján teszünk fel ellenörző kérdéseket egy beugró ZH keretében (10 perc) .

A mérés háttere:

A genotipizálási mérés során egyszerre sok egyed DNS-én tudjuk megvizsgálni az egyes helyeken levő allélokat. Először a begyűjtött mintákból kivonjuk a DNS-t, majd a számunkra érdekes területekről (A single nucleotide polymorphism -SNP-k – azaz az egynukleotidos polimorfizmusok 150 bázispárnyi környezetéről) másolatokat szaporítunk polimeráz láncreakcióval (PCR).

A vad és mutáns allélokhoz tartozó DNS láncokat különböző színű fluoreszcens festékkel festjük meg. Ezután a mintákat egy olyan sziliciumlapra helyezzük, amelyen a komplementer DNS láncok helyezkednek el. A szaporított minták ezekhez kötődve bizonyos hullámhosszok alatt fluoreszkálnak. Ekkor készítünk a két színcsatorna alatt egy-egy felvételt, majd a későbbiekben részletezett képfeldolgozási eljárással megfigyeljük az egyes pontok fényességét, valamint a pontok további jellemzőit is rögzítjük.

Ezután az egy SNP-hez tartozó mintákat összegyűjtjük, és egy diagrammon ábrázoljuk. A diagramm X tengelye a minta színaránya. Ha a pont teljesen kék, akkor jobbra, ha teljesen zöld akkor pedig balra kerül. Az Y tengely pedig a pontok összegzett intenzitása.



A mintapontok naív klaszterezése adja meg a genotípust, de ez nem felel meg teljesen a valós genotípusnak, a mérés megbízhatatlansága miatt. Egy jelentősen drágább módszerrel történő validáció során pontosan megállapíthatjuk az egyedek genotípusát. A validációt felhasználva szeretnénk megállapítani hogy milyen paraméterek mellett valószínű hogy a mérés hibás genotípust fog az egyedhez hozzárendelni.

A megoldások minőségét az ROC (receiver operating characteristics) görbével tudjuk szemléletesen ábrázolni. A görbe a megoldás érzékenységét veti össze a specifikusságával.

Ellenőrző kérdések:

- Mi az a PCA?
- Milyen tengelyek szerepelnek az ROC görbén?
- Mire használjuk a mérés kezdeti fázisában a PCR-t?
- Milyen jellegü zavart okozhatnak a rendkívül zajos paraméterek egy MLP tanításában ha kevés tanító pontunk van és nagy az adatdimenzió?

Házi Feladat:

Készítsen Matlab környezethez egy függvényt amely képes kiszámítani egy ROC görbe alatti területet!

A mérés menete:

A mérés során a Matlab neural network és statistics toolbox-ait (erről a mérésvezető rövid ismertetést ad igény szerint a mérés elején) fogjuk használni. Érdemes már otthon is, ha van lehetősége rá, megismerkedni a neural network toolbox-al.

Az adatmátrix egy sora egy adott bemeneti minta. Az előkészített adatok letölthetők a tárgy honlapjáról a BIR_1meres_data.csv állományból

(http://www.mit.bme.hu/oktatas/targyak/vimim306/feladat), a megértést segíti a segédletben adott paramétermagyarázat. Töltse le, és ismerkedjen meg az adatokkal!

A mérés során jegyzőkönyvet kell készíteni. A jegyzőkönyv lényegre törő, világos kell legyen. Minden feladnál tartalmaznia kell a feladat megfogalmazását, a megoldás menetének főbb részleteit, az eredményeket és az eredmények értékelését.

1. feladat

A. Importálja az adat fájlt, és válassza külön a későbbi bemeneti és célváltozókat a segédletben megadott információk alapján! A célváltozó legyen a Clustering_error_bool, míg a bemeneti változók legyenek a tőle jobbra levő oszlopok értékei. A matlab/file menu/import paranccsal lehet betolteni (next next finish). Ügyeljen arra, hogy a Matlab importálásnál szétválasztja a csv fájlt numerikus és szöveges részekre!

B. Vizsgálja meg PCA-val a bemeneti adatok legnagyobb sajátértékeit! Figyelje meg hogy mennyi sajátvektorral tudja a jól (>99%) kifeszíteni a mintahalmaz varianciáját! Használja a *princomp()* függvényt! Használja a `help princomp` parancsot, ha elakad az adatok értelmezésében!

```
[COEFF,SCORE,latent] = princomp(X)
```

C. Dimenzióredukció: transzformálja az az adatait az 1-5 legnagyobb variancát megőrző sajátvektorok terébe, valamint a 3 legjobb és két véletlenszerűen választott sajátvektor terébe!

2. feladat

A. Vizsgálja meg meg a két kitüntetett sajátvektor definiálta biplotban a változókat és az adatokat a három legjobb sajátvektor terében! (természetesen azért csak három mert ennél több dimenziót nehézkes lenne vizualizálni) Lát-e valamilyen szeparációt a pontok között? Ha igen, készítsen róla képernyőmentést.

```
biplot(coeffs(:,1:3),'scores',score(:,1:3))
```

3. feladat

A. Tanítson egy bináris kimenetű osztályozó perceptront (az `nntool` eszközzel) az ismert validáció alapján (Clustering_error_bool) a legjobb PCA származtatott értékeket felhasználva! Értékelje a kapott hálózat eredményét! Ne felejtse transzponálni az adatait, hogy a bemeneti minták oszlopokban helyezkedjenek el.

B. Tanítson egy egy rejtett rétegű, maximálisan bemenetek száma neuronnal rendelkező MLP-t a hiba predikciójára az ismert validáció (klaszterezési hiba), valamint az előzőekben kiválasztott paraméterek alapján. Ez lesz a visszautasítási modellje, amelyet tovább kell vizsgálnia. Készítsen MLP-t a teljes bemeneti adathalmazzal, valamint az 1-5 legjobb származtattot PCA paraméterrel és a 3 legjobb és 2 véletlenszerűen választott paraméterrel is!

4. feladat

A. Vizsgálja az előző feladatban elkészített modellek specifikusságának és érzékenységének alakulását egy elutasítási küszöb paraméter segítségével! Ábrázolja az ROC görbét a *roc* és a *plotroc* függvények segítségével!

```
[tpr,fpr,thresholds] = roc(targets,outputs)
```

B. Számolja ki a görbe alatti területeket és értékelje az egyes modellek teljesítményét a házi feladatként elkészített ROC görbe alatti területet kiszámoló függvénye segítségével!

5. feladat

Válassza ki a szerinte legjobban teljesítő modellt (választását indokolja) és számítsa ki a hibamodell költségét az alábbi költségmátrix alapján, valamint rajzoljon fel súlyozott hasznossági görbét is a mátrix alapján! Ehhez írjon egy függvényt amely az alábbi módon súlyozza a hibát:

Költségmátrix:

Taqman/RawRead	AA (0)	AG (1)	GG (-1)	Eldobási költség
AA	0	1	2	(Sum-SumAA)/Sum = 0.394
AG	1	0	1	(Sum-SumAG)/Sum=0.658
GG	2	1	0	(Sum-SumGG)/Sum =0.9343

A mátrix 3x3 as része azt adja meg, hogy milyen költséget rendelünk ahhoz, hogy egy genotípust hibásan adunk meg. Látható, hogy egy klasztert tévedni 1-be kerül, míg két klaszterrel arébb sorolni a mintát jóval drágább (AA helyett GG).

A mátrix utólsó oszlopa pedig az eldobás költségét adja meg, amely értelemszerűen alacsonyabb, mint a hibás válasz adása, és ára attól függ, hogy hány darab olyan mintánk volt az egész mérésben. Értelemszerűen a ritkább halmazba tartozó mérés eldobása drágább.

6. feladat (Bónusz feladat)

Vizsgálja meg az Asztma es a genotipus (Taqman) korrelaciojat!

1. Introduction to genotyping.

The distinct stages of the genotyping measurement are followed by complex post processing. First we apply digital image processing algorithms to the raw image information to produce real parameter values, which are then clustered to produce the nominal results of the measurement.

The genomic SNP Core Facility laboratory at the Department of Genetics Cell- and Immunobiology of the Semmelweis University also has to face the same problems as mentioned before. Beside commercial measurements the laboratory investigates the genetic backgrounds of various diseases such as asthma and allergy. In the course of their work, the researchers want to explore the genetic causes of the diseases, therefore it is essential to define the SNPs associated with the illness.

2. Terminology

2.1. Genome

The genome contains the entire hereditary information of living organisms, which in most cases is coded by DNA. The expression was created by merging the words gene and chromosome by Hans Winkler, botanics Professor of the University of Hamburg in 1920.

DNA is a nucleic acid with the shape of a double spiral that is created by the connections of four nucleotide base pairs; adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T).

The spiral is held together by the complementary bases on each strand. Adenine and thymine and are connected by double hydrogen bonds, while cytosine and guanine are connected by triple hydrogen bonds. DNA strands in the genome are organized into chromosomes, and are present in the human body in the form of two homologous chromosomes, forming a chromosome pair. In human cells there are 23 pairs of chromosomes, each member of the pairs originates from one of the parents. The corresponding loci of the homologous chromosome pairs are called alleles, thus the human chromosomes are usually biallelic. The phenotype is any observable characteristic or trait of an organism; such as its morphology, biochemical or physical properties or behavior.

2.2. Genotype

The genotype is the genetic trait that cannot be directly observed. It is the specific genetic sequence of a cell, and organism, or an individual, i.e. the specific allele make up of the individual.

Genotyping is the process of determining the genotype of an individual by the use of biological assays. It provides measurement of the genetic variation between members of a species.

2.3. Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNP, pronounced ‘snip’) are the most common type of genetic variation. A SNP is a single base pair mutation at a specific locus, usually

consisting of two alleles (where the rare allele frequency is ≥ 1%). SNPs are often found to be the biomarkers of many human diseases and are becoming of particular interest in pharmacogenetics.

A SNP is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles : C and T. Almost all common SNPs have only two alleles.

Within a population, SNPs can be assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another.

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. However, their greatest importance in biomedical research is for comparing regions of the genome between cohorts (such as with matched cohorts with and without a disease

2.4. Genotyping methods and the biochemistry of genotyping

2.4.1 Beckman Coulter's GenomeLab SNPstream Genotyping System

The GenomeLab SNPstream Genotyping System utilizes a proprietary method called SNP Identification Technology for the detection of single nucleotide polymorphisms (SNPs). SNP Identification Technology is a non-radioactive, single-base primer extension method that can be performed in a variety of formats. It relies upon the ability of DNA polymerase to incorporate dye labeled terminators to distinguish genotypes.

2.4.2. Probe/Tag Technology

The SNP Identification Technology method is informative because it provides direct determination of the variant nucleotides. SNP Identification Technology also provides significant research accuracy to genotyping because it incorporates — after PCR — a two-tiered detection utilizing base-specific extension by polymerase followed by hybridization-capture. This two-tiered detection step ensures accurate and highly discriminant analysis.

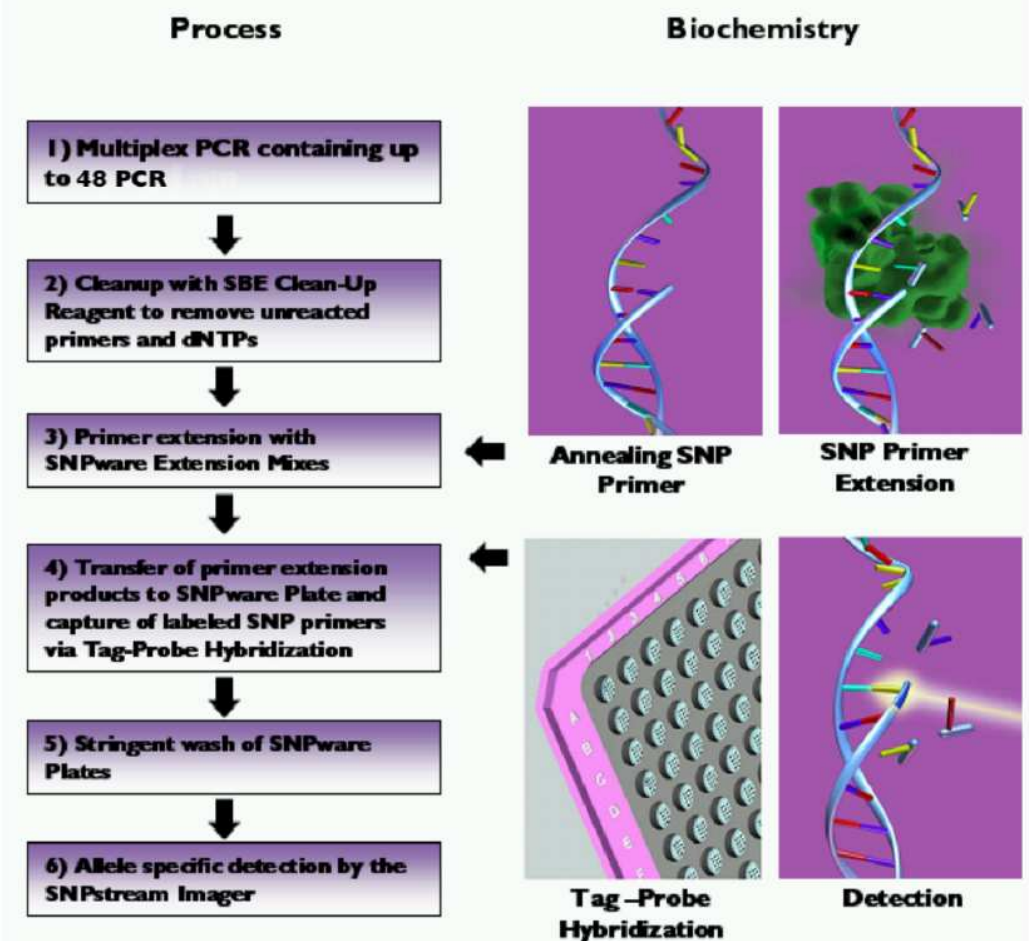
The hybridization capture step utilizes a tag-probe approach. The SNP Identification Technology primer is a single strand DNA containing a template specific sequence appended to a 5' non-template specific sequence. Tag refers to the sequence attached to the SNP Identification Technology primer that is captured by specific probe bound to glass surface. The probe refers to a unique DNA sequence attached to the glass surface of every well in a 384 tag-array plate that specifically hybridizes to one tag. The probes bound covalently to the glass surface enable the interrogation of up to 12-plexed or 48-plexed nucleic acid reaction products. The SNP reaction product into which the tag

has been incorporated will hybridize to the corresponding probe bound covalently to the glass surface.

2.4.3. SNP Assay

SNP biochemistry for the GenomeLab SNPstream Genotyping System involves the following steps, as shown schematically after multiplex PCR amplification, amplicons containing the SNP of interest (step 1), unincorporated nucleotides and primers are removed enzymatically (step 2). In step 3, extension mix and a pool of tagged SNPware primers are added to the treated PCR.

SNPware primers hybridize to specific amplicons in the multiplex reaction, one base 3' to the SNP sites. The tagged primers are extended in a two-dye system, by incorporation of a fluorescent labeled chain terminating acyclonucleotide. Two-color detection allows determination of the genotype by comparing signals from the two fluorescent dyes.



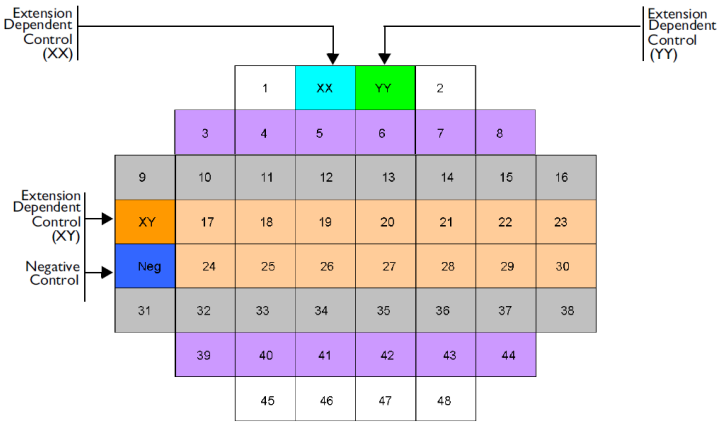
SNP Identification technology for SNPStream

The extended SNPware primers are then specifically hybridized to unique probes arrayed in each well. The arrayed probes capture the extended products (step 4) and allow for the detection of each SNP allele signal (step 5). Stringent washes remove free dye-terminators and DNA not hybridized to specific probes.

2.4.4. Control spots

Two self-extending control oligonucleotides are included in each extension master mix and are extended with either the blue or green dye-labeled terminator during the primer extension thermal cycling.

The array of capture oligonucleotides attached to the glass surface in each well of a 384-well plate includes three positive controls and one negative control. The XY control spot is a heterozygous control which has a mixture of two capture probes that allow hybridization of both blue and green control oligonucleotides. The XX control spot has a capture probe that allows hybridization of the blue control oligonucleotide. The YY control spot has a capture probe that allows hybridization of the green control oligonucleotide.

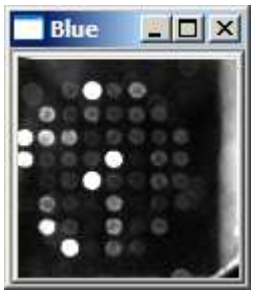


Control spots and spot layout for 48plex plates

3. Image processing overview and image based quality parameters

3.1 Noise patterns

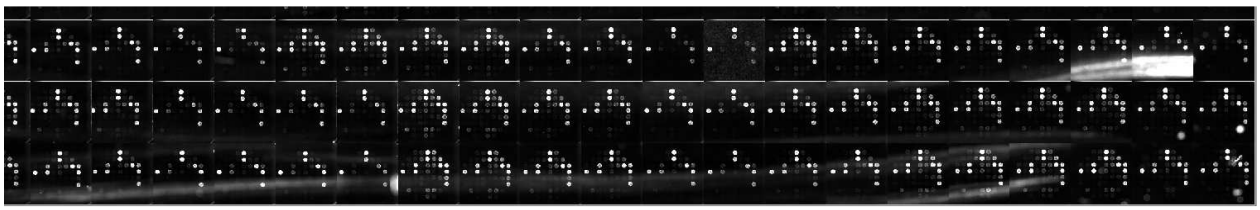
Calculating the intensity is not merely enough to obtain reliable genotyping data from the scanned images, since many artifacts and errors can distort the scanned image, such as those shown below



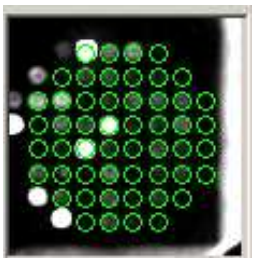
Noisy images

Specs of dust on the plate

Marks left by residual chemicals from the PCR reaction



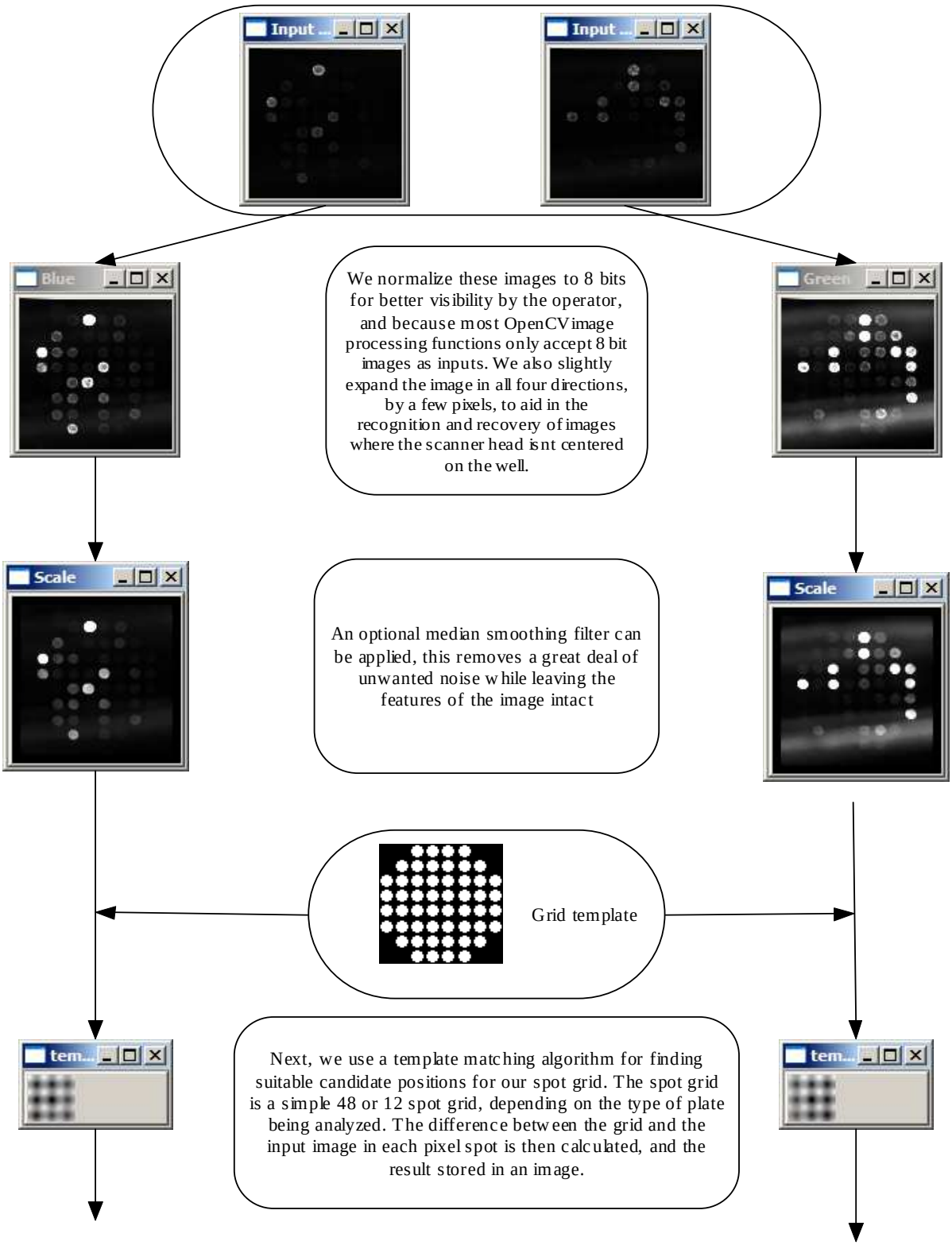
Banding caused by improper wiping of a plate before scanning (very rare but can be severe)

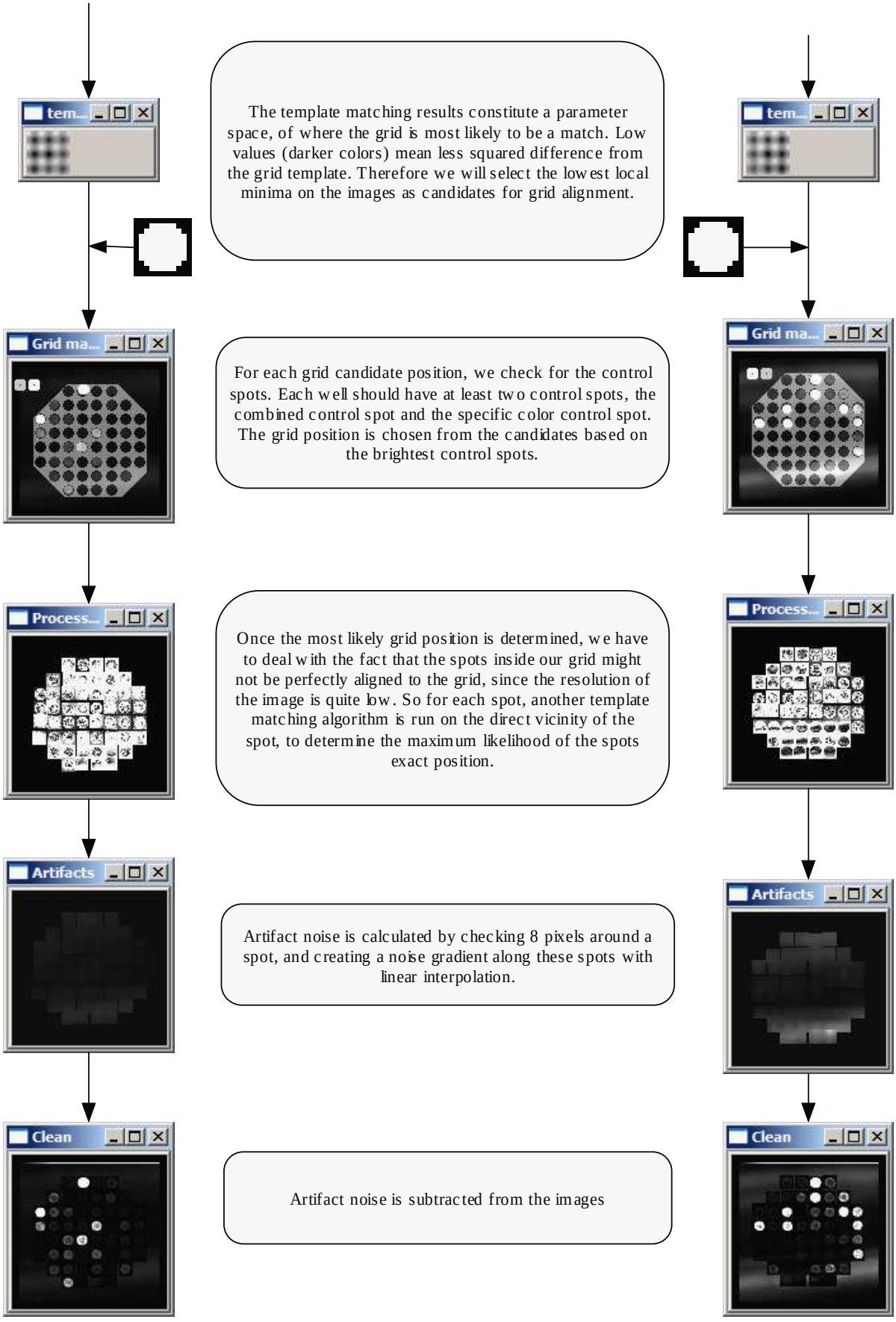


The scanner is not centered above the well, so some spots may be partially missing as well as the measurements from the whole well being false.

3.2. Image Processing Overview

First, we read the corresponding 16 bit raw images for each well:




tem...
The template matching results constitute a parameter space, of where the grid is most likely to be a match. Low values (darker colors) mean less squared difference from the grid template. Therefore we will select the lowest local minima on the images as candidates for grid alignment.
tem...
Grid ma...
For each grid candidate position, we check for the control spots. Each well should have at least two control spots, the combined control spot and the specific color control spot. The grid position is chosen from the candidates based on the brightest control spots.
Grid ma...
Process...
Once the most likely grid position is determined, we have to deal with the fact that the spots inside our grid might not be perfectly aligned to the grid, since the resolution of the image is quite low. So for each spot, another template matching algorithm is run on the direct vicinity of the spot, to determine the maximum likelihood of the spots exact position.
Process...
Artifacts
Artifact noise is calculated by checking 8 pixels around a spot, and creating a noise gradient along these spots with linear interpolation.
Artifacts
Clean
Artifact noise is subtracted from the images
Clean

4. Data file format

plate_ID	Shows which plate the sample originated from, 2 plates in total
well_ID	Shows which well the sample was in, wells are marked from a01 to p24
sample_ID	Identifier of the sample, eg which person the DNA came from
Asthma	Shows wether the patient had Asthma or not (1=true)

Target variables

TaqMan	Results of validation
Rawread_HW_clustering	Our genotype calls (inaccurate, based on a greedy clustering method
Taqman_int	Same as previous, but in integer form (-1= AA, 0=AG, 1=GG)
Rawread_int	
Clustering_error_int	Clustering error abs(target variable- our result)
Clustering_error_bool	Boolean clustering error

Input variables

totalnoiseb	Total amount of noise in the blue channel, as calculated from the artifact supression
snr_b	Signal to noise ratio in blue channel
totalnoiseg	Total amount of noise in the green channel, as calculated from the artifact supression
snr_g	Signal to noise ratio in blue channel
certainty	Distance from cluster center
certaintygradient	Weighted distance from cluster center (lower intensity is punished more)
intensity_B+G	Total sum of the intensity of the blue and green spots
intensity_b/b+g	Ratio of blue to green channel brightness, where 1=all blue and = all green
logavgb	Average blue intensity across the spot.
logintensityb	Blue channel intensity sum
logintensity varianceb	Intensity variance over the spot (see processing flowchart)
logcircularityb	Metric of circulatiry, low values mean uniform circles, high values are more amorphous
logbasenoiseb	Base noise level for the spot
logArtifact corrected intensityb	Artifact suppressed intensity rating, this value is used to calculate the clustering plots**.**

logcorner1b logcorner2b logcorner3b logcorner4b logcorner5b logcorner6b logcorner7b logcorner8b logcenterb Intensity values on the 8 corners of the spot image, and the intensity of the images in the center point. Note: Be wary when

using these parameters, as they do not represent too much useful information with relation to the genotype or the error of the sample. They can be regarded as quasi random variables.

logavgg logintensityg logintensity varianceg logcircularityg logbasenoiseg logArtifact corrected intensityg logcorner1g logcorner2g logcorner3g logcorner4g logcorner5 logcorner6g logcorner7g logcorner8g logcenterg

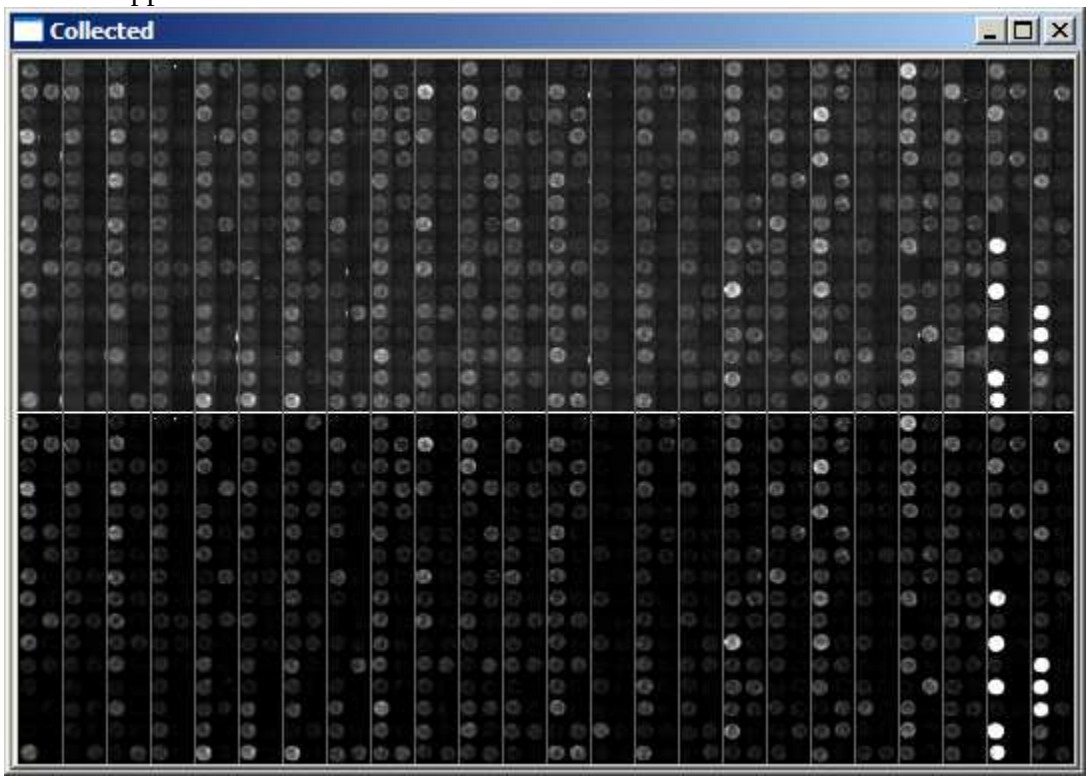
Same as previous, but for the green channel.

6. Appendices:

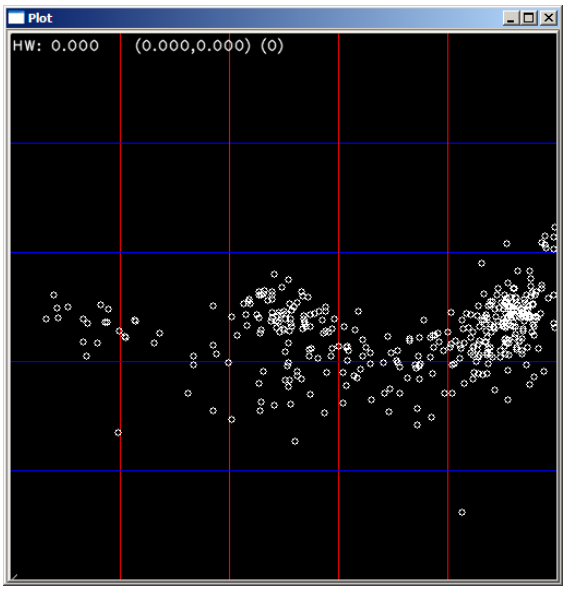
Plate number: 0000501603

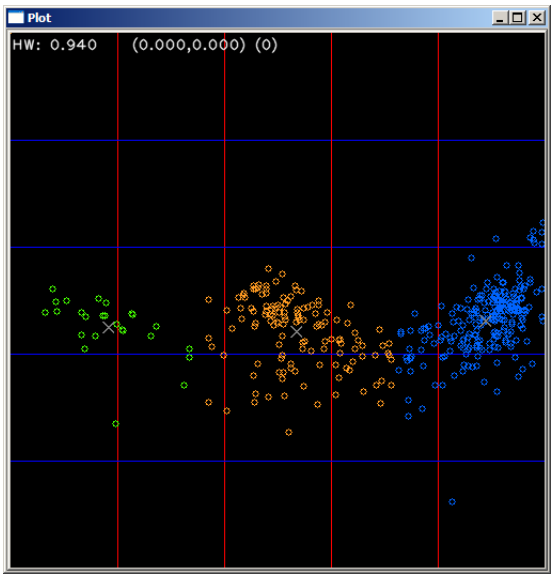
Collected view for SNP RS7167

Each collected sample for this SNP is shown on the image below. Blue and green spots are placed next to each other while the bottom half of the image shows the version with artifact suppression.



Unclustered plot:





K-means clustering:

Blue view for plate:

Green view for plate:

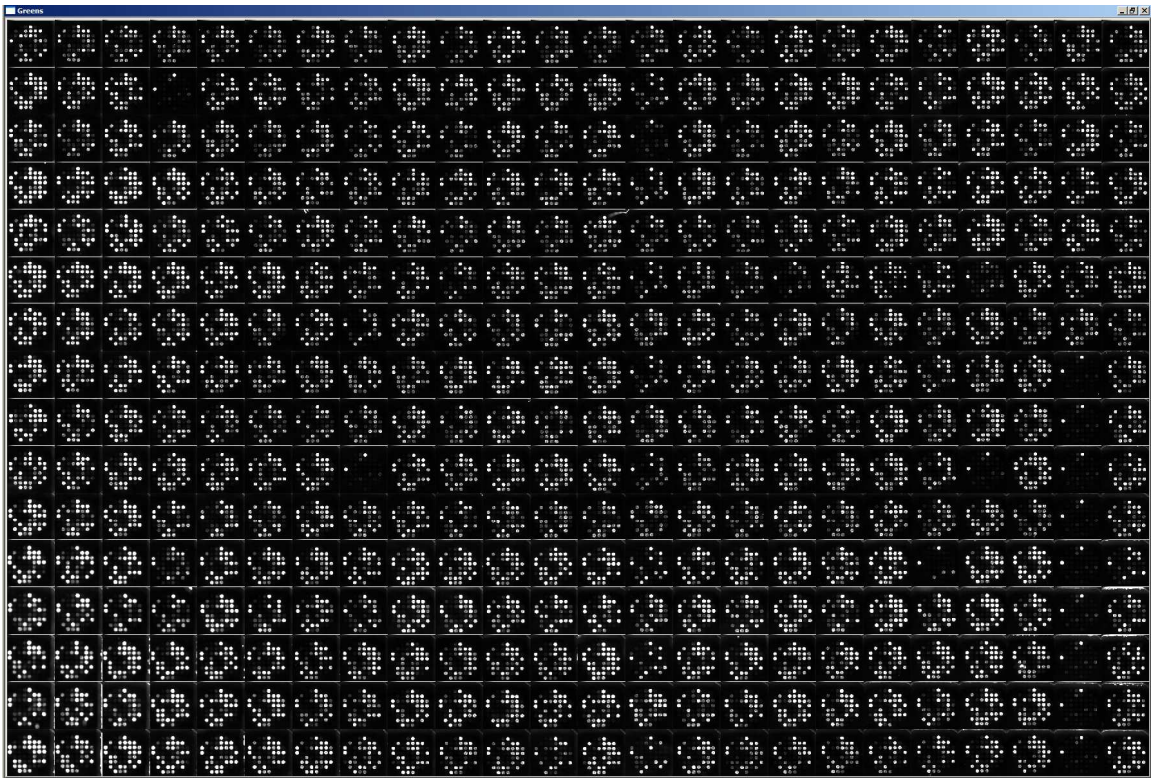
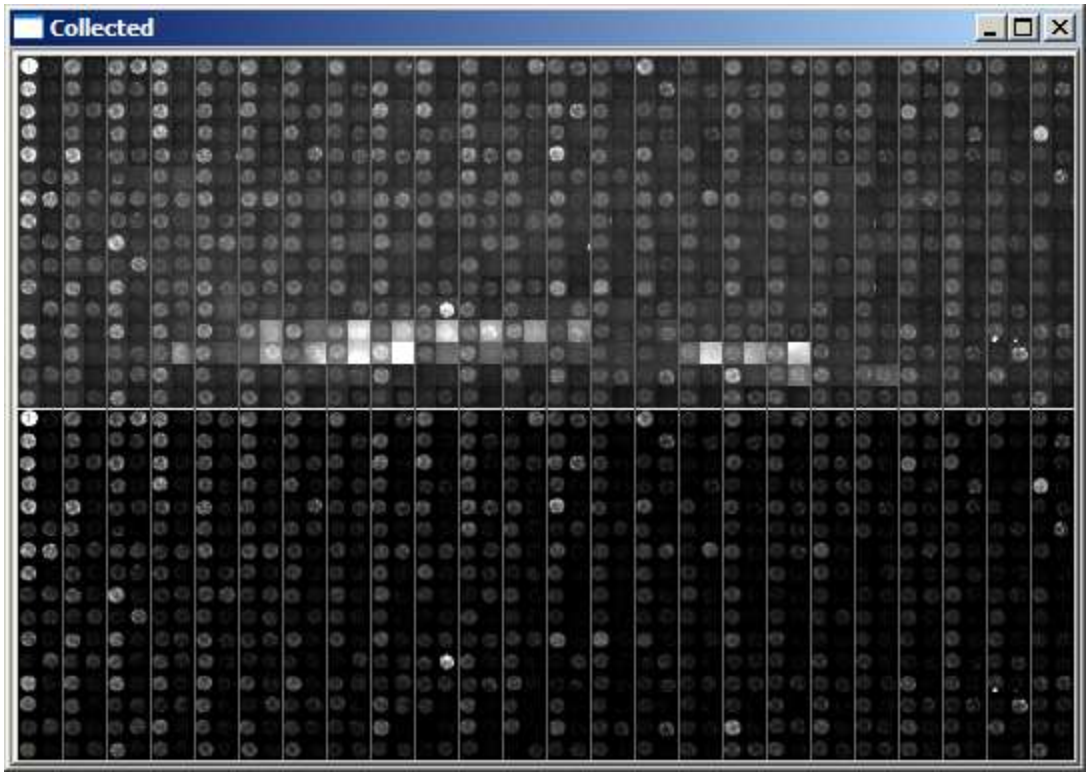
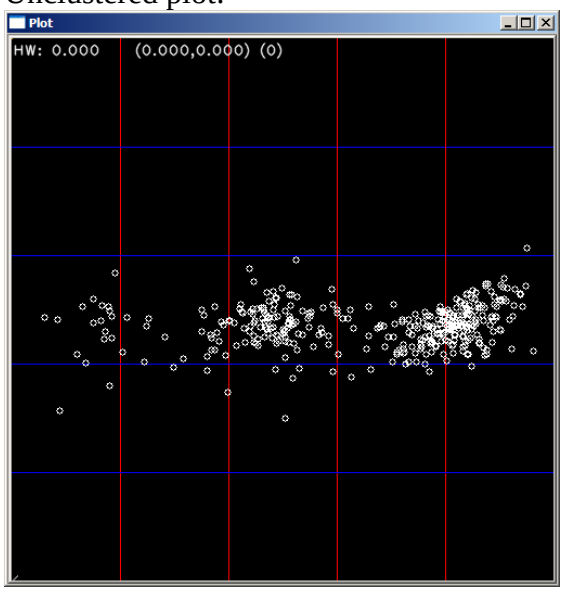


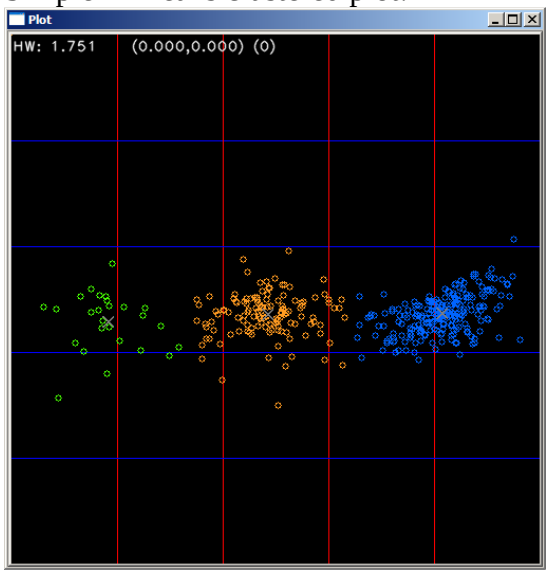
Plate number: 0000508403:
Collected view:



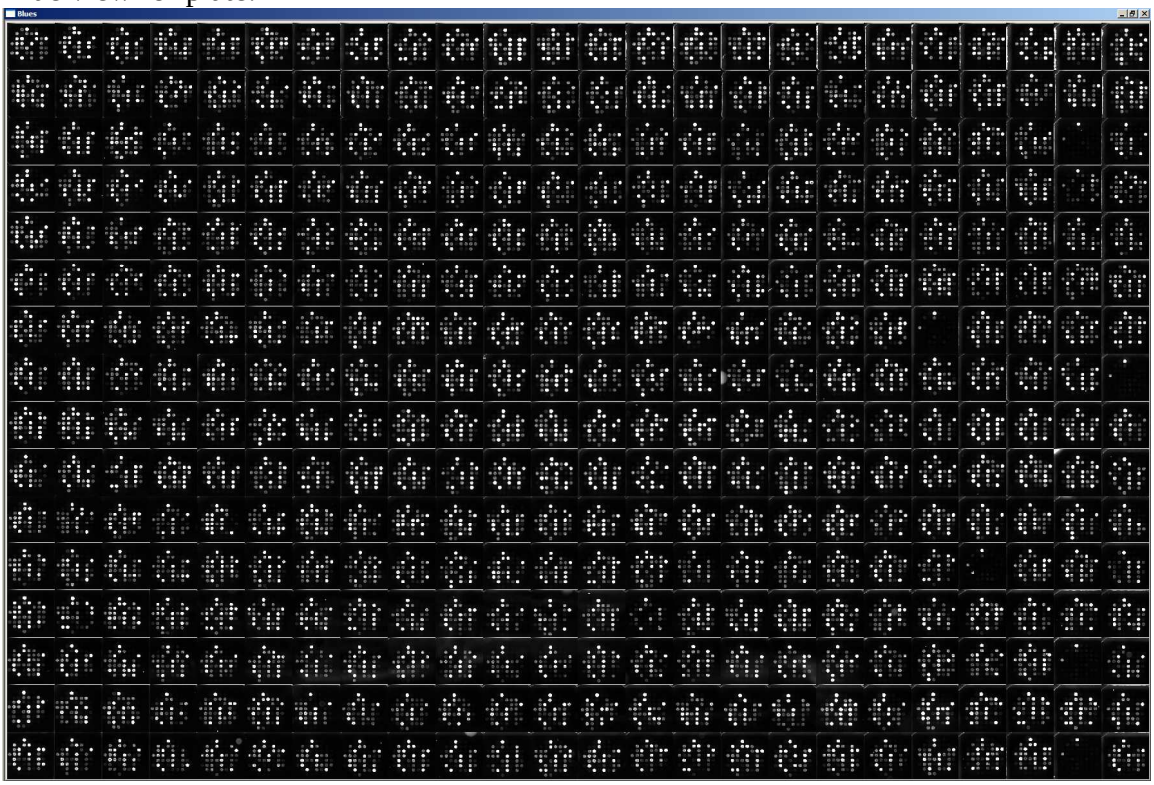
Unclustered plot:



Simple K-means clustered plot:



Blue view for plate:



Green view for plate: