Illumina's Genotyping Data Normalization Methods

I. Abstract

Many of Illumina's customers have asked for guidance in exploring normalization procedures for Illumina's raw genotyping data. This purpose of this document is to provide general guidelines for those researchers who wish to try their own normalization procedures, and to detail Illumina's standard normalization method. Regardless of the normalization algorithm used, it is necessary to apply such an algorithm by sub-bead pools as defined below. Illumina's standard normalization algorithm is implemented as the first step in SNP genotyping data analysis. The intensity data are normalized automatically when they are loaded into Illumina's BeadStudio software.

II. Introduction

Any normalization procedure applied to Illumina's genotyping data must be applied on the sub-bead pool level. A sub-bead pool is a set of beads that were manufactured together and are located in roughly the same analytical location (stripe) on a BeadChip. The sub-bead pool information for the Human-1, HumanHap300, Hap550, and Hap650Y can be obtained by contacting Illumina's technical support (techsupport@illumina.com) and is provided as a bead pool manifest (.bpm) file. To obtain this data for future product releases, please contact technical support. As new products become available, follow the same process to obtain the sub-bead pool information and raw data.

Because the performance of external controls can vary from sample to sample, Illumina's standard normalization is performed without the use of external controls. Illumina has developed a self-normalization algorithm which draws on information contained in the array itself. This approach contributes to the generation of high-quality, accurate genotyping calls. You can use the procedures described in this document to replicate the steps typically performed by Illumina's BeadStudio software (the BeadStudio Genotyping Module) to convert raw X and Y (allele A and allele B) signal intensities to normalized values. Normalized values are always used to analyze standard genotyping calls, Loss of Heterozygosity (LOH), and Copy Number (CN).

The normalization algorithm is designed to adjust for channel-dependent background and global intensity differences, and to scale the data. It is important to note that the normalization process uses the information that links a bead type to a sub-bead pool. Typically, a BeadSetID (represented by a unique identifier) corresponds to the content represented on an individual stripe on a BeadChip. However, the normalization process takes place on the sub-bead pool level, not on the stripe level. There are 24 sub-bead pools for the Human-1, 14 for the Hap240S, 11 for the Hap300, 25 for the Hap550, and 27 for the Hap650Y.

The sub-bead pool information for Human-1, HumanHap300, HumanHap550, and HumanHap650Y BeadChips is provided as a supplementary file that links the ID of each SNP locus to its representative bead pool. To determine which BeadSetID contains a SNP locus of interest, you can search this table with a VLOOKUP function. For details about each sub-bead pool, please refer to Section VIII of this document.

III. Estimating normalization parameters

Illumina uses a 6-degree of freedom affine transformation to normalize sample intensities. The six parameters are offset_x, offset_y, theta, shear, scale_x, and scale_y. The normalization process consists of five main steps:

1. Outlier removal
2. Background estimation (offset_x, offset_y)
3. Rotational estimation (theta)
4. Shear estimation (shear)
5. Scaling estimation (scale_x, scale_y)

Figure 1 depicts the stages of the normalization process.

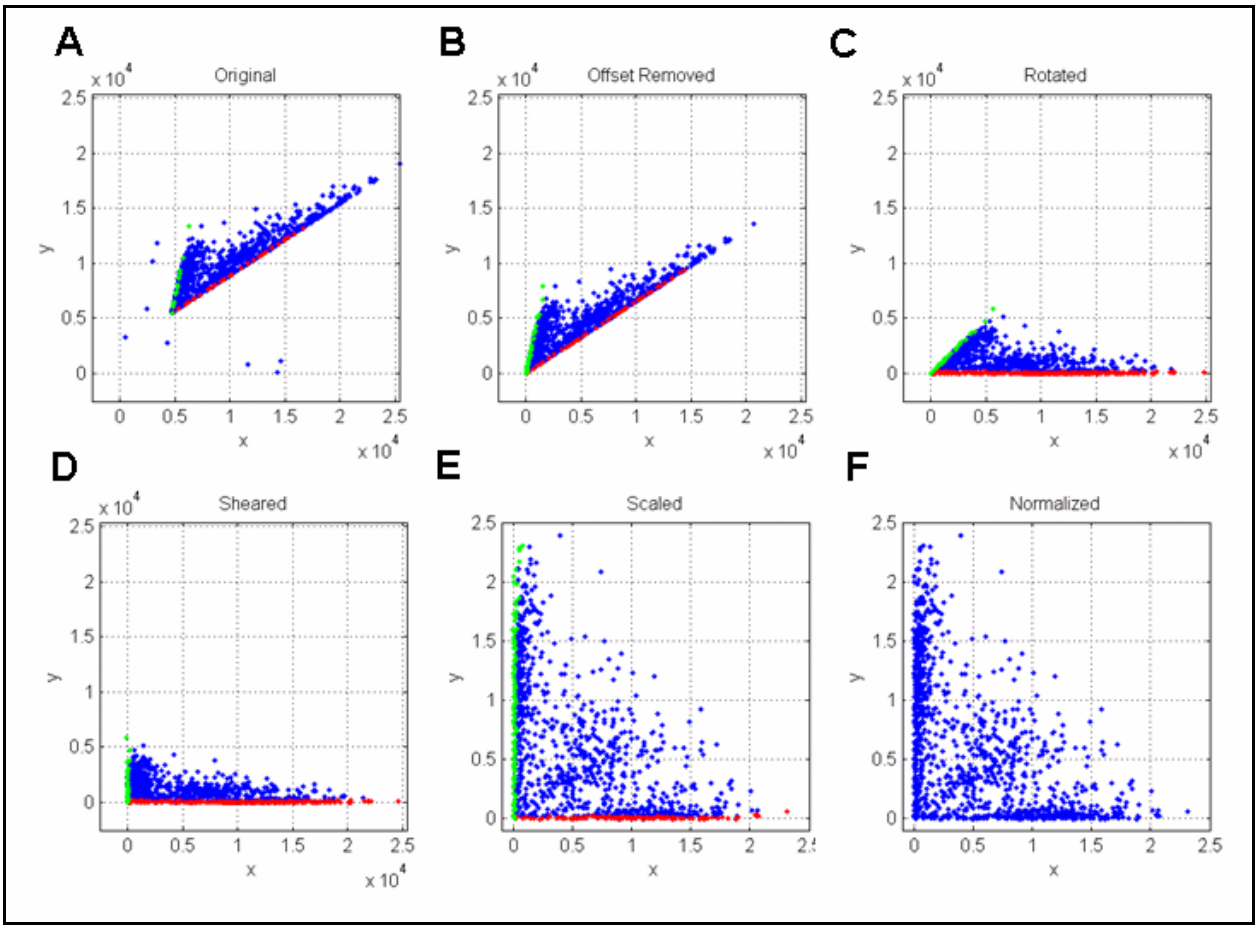

Figure 1: Graphical Representation of the Process Used to Normalize Genotyping Data

1. Outliers (Figure 1-A)

   Outlier SNPs are removed from consideration during normalization parameter estimation. These SNPs are only considered outliers during the normalization process and are not excluded from downstream analysis. A SNP is considered an outlier if its intensity meets any of the following criteria:

   - Its value of x, y, or x/(x+y) is smaller than either the 5th smallest or the 1st percentile (whichever is smaller) of those values across all SNPs.
   - Its value of x, y, or x/(x+y) is larger than either the 5th largest or the 99th percentile (whichever is larger) of those values across all SNPs.

2. Translation (Figure 1-B)

   a. An x-sweep is performed by sampling 400 points along the x-axis, from the smallest x value to the largest. The closest SNP to each sampled point along the axis is added to the set of candidate homozygote As.
   b. The same analysis is performed along the y-axis to find the candidate homozygote Bs.
   c. A straight line is fit into candidate homozygote A alleles.
   d. A straight line is fit into candidate homozygote B alleles.
   e. The intercept of the two lines is computed, and this coordinate corresponds to offset_x and offset_y.

3. Rotation (Figure 1-C)

   a. The points are corrected for translation and another x-sweep is performed to determine a set of control points.
   b. A straight line is fit into the control points. The angle between this line and the x-axis defines the amount of rotation in the data. This angle corresponds to the theta parameter.

4. Shear (Figure 1-D)

   a. The points are corrected for rotation and another y-sweep is performed to determine a set of control points.
   b. A straight line is fit to these control points. The angle of this line identifies the shear parameter.

5. Scale (Figure 1-E)

   a. The points are corrected for shear, and another x-sweep is performed to identify a set of virtual points.
   b. A statistical robust measure of the mean of these control points is used to determine scale_x.
   c. A Y-sweep is done, and some virtual points are identified via triangulation. A statistical robust measure of the mean of these control points is used to determine scale_y.

6. Final Results (Figure 1-F)

   Figure 1-F depicts the final set of normalized data points.

IV. Performing the normalization

To convert raw coordinates (x raw and y raw) to normalized coordinates (x normalized and y normalized), perform the following operations for each SNP, using the normalization parameters determined for that SNP's sub-bead pool:

1. temp x = xraw - offset_x

   temp y = yraw - offset_y

2. temp x2 = cos(theta) * temp x + sin(theta) * temp y

   temp y2 = -sin(theta) * temp x + cos(theta) * temp y

3. temp x3 = temp x2 - shear * temp y2

   temp y3 = temp y2

4. x n = temp x3 / scale_x

   y n = temp y3 / scale_y

V. Important facts to remember when performing your own normalization process

Illumina's normalization process must take place on a sub-bead pool level. This holds true regardless of the normalization process used. If you intend to use your own custom normalization process and not the process described here, it still must occur on a sub-bead pool level. Not incorporating this data will result in the generation of unsatisfactory and unrepresented data. Use the beadset-lookup number for each SNP to identify its bead pool.

VI. Plotting and Visualizing Data

To visualize the data after normalization, the genotyping data are transformed to a polar coordinate plot of normalized intensity $R = X_{norm} + Y_{norm}$ and allelic composition (copy angle), using the equation theta = (2/pi)*arctan2 (Y_{norm},X_{norm}), where X_{norm} and Y_{norm} represent transformed normalized signals from alleles A and B for a particular locus.

VII. Concluding remarks

Illumina's genotyping data require normalization in order to be as canonical as possible. This process helps generate precise, accurate, high-quality genotyping calls. Self-normalization uses information contained within the array itself (BeadSetID) plus five essential steps including outlier removal, background estimation, rotational estimation, shear estimation, and scaling estimation. When working with unnormalized, raw genotyping data (X raw and Y raw signal intensities), use the aforementioned protocol as a guideline for your own analyses. If a custom normalization procedure is used, be sure to apply it at a sub-bead pool level; otherwise, data quality will be severely compromised and may yield inaccurate conclusions.

VIII. Supplementary Information

The following tables provide detailed information about the content of each BeadChip and its corresponding BeadSetIDs.

Number of Sub-Bead Pools per Product

Product	# Sub-Bead Pools
Human-1	24
Hap300	11
Hap240S	14
Hap550	25
Hap650Y	27

BeadSetIDs for Human-1

Product Content	Stripe	Pool	BeadSetID in BPM[1]
Human-1	1	1a	1138231287
Human-1	1	1b	1148060415
Human-1	2	2a	1116886462
Human-1	2	2b	1120911303
Human-1	3	3a	1136944926
Human-1	3	3b	1148229933
Human-1	4	4a	1137037017
Human-1	4	4b	1148282830
Human-1	5	5a	1137093177
Human-1	5	5b	1148330242
Human-1	6	6a	1137143328
Human-1	6	6b	1148382903
Human-1	7	7a	1145741357
Human-1	7	7b	1148429148
Human-1	8	8a	1115759993
Human-1	8	8b	1120739624
Human-1	9	9a	1117936208
Human-1	9	9b	1120819191
Human-1	10	10a	1118515762
Human-1	10	10b	1120866864
Human-1	11	11a	1119248973
Human-1	11	11b	1121106758
Human-1	12	12a	1119828527
Human-1	12	12b	1121160224

BeadSetIDs for Hap300, 240S, 650Y

Product Content	Stripe	Pool	BeadSetID in BPM
Hap300	1	1a	1136666071
Hap300	1	1b	1136680268
240S	1	1c	1138477660
240S	1	1d	1140177724
Hap300	2	2a	1135221993
240S	2	2b	1138831736
240S	2	2c	1140283518
Hap300	3	3a	1135451484
240S	3	3b	1138865330
Hap300	4	4a	1135655706
240S	4	4b	1139058074
Hap300	5	5a	1135859225
240S	5	5b	1139251008
240S	5	5c	1142813120
Hap300	6	6a	1136063447
240S	6	6b	1142733151
Hap300	7	7a	1136266966
240S	7	7b	1139636193

[1] Bead Pool Manifest

Product Content	Stripe	Pool	BeadSetID in BPM
Hap300	8	8a	1136868904
240S	8	8b	1139827361
240S	8	8c	1141148388
Hap300	9	9a	1137495746
240S	9	9b	1141367091
Hap300	10	10a	1137813382
240S	10	10b	1141795769
Hap650Y	11	11	1145224029
Hap650Y	12	12	1145605848
Note: 240S + Hap300 = Hap550			

IX. Patent Protection Notice

All of the processes described in this document are protected by patent U.S. No. 7,035,740.

For Research Use Only

Pub. No. 970-2006-010 26Sep06