TESTING A METHODOLOGY FOR IDENTIFYING CLUSTERED ALLELE LOSS USING SNP ARRAY DATA

by

Ping Zheng

Bachelor of Medicine, Fujian Medical College in P. R. China, 1985

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

By

Ping Zheng

It was defended on

December 5, 2007

and approved

by

Thesis Advisor:
Richard Day, PhD
Associate Professor, Biostatistics
Assistant Professor, Infectious Diseases & Microbiology
Graduate School of Public Health
Assistant Professor, Anthropology
School of Arts and Sciences
University of Pittsburgh

Committee Member:
Stephen Grant, PhD
Associate Professor, Environmental and Occupational Health
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Chien-Cheng (George) Tseng, ScD
Assistant Professor, Biostatistics
Graduate School of Public Health
University of Pittsburgh

Richard Day, PhD

TESTING A METHODOLOGY FOR IDENTIFYING CLUSTERED ALLELE LOSS USING SNP ARRAY DATA

Ping Zheng, MS

ABSTRACT

HumanHap550 Genotyping BeadChip provides a platform allowing for genotyping of single nucleotide polymorphisms (SNPs) greater than 550,000 loci. Such SNPs genotyping array technology makes it possible to identify genetic variation in individuals and across populations, profiling somatic mutations in cancer and loss of heterozygosity (LOH) events, amplifying deletions of regions of DNA, as well as possibly evaluating germline mutations in individuals. This study particularly focuses on analysis of clusters of Mendelian inconsistencies (MIs) in the SNPs array for six Russian radiation worker family trios, in order to identify the type of deletion variants for offspring such as inherited parental deletion variants (PDVs), spontaneous mutations (SMs) and germline mutations (GMs). By adapting the Bayesian theorem combining with the hereditary rule, this study presents an exciting result because 96.15% of genotypes in six selected clusters under the investigation could be identified as either PDVs or SMs/GMs, with two clusters are perfectly identified as SMs/GMs. This opens an avenue for further investigation of whether external environmental exposures (e.g., ionizing radiation) can effect the frequency of deletion variants (i.e., germline mutations) occurring in the offspring of highly exposed nuclear workers. While the applied methodology provides a practical means to recognize the genomic

variations within the SNPs array some weaknesses of the study have been observed; particularly, the control group which consists of 112 individuals of Yoruba, Han Chinese, Japanese and Mormons is of deficiency on its sample size, diverse ethnicity and DNA process compared to the case group, and unclean potential hemizygous SNPs (i.e., Mendelian inconsistencies). Further statistical investigation and research needs to be conducted in order to overcome the weaknesses observed in the study; hence, the methodology introduced would be further of enhancement in its reliability and validity and it should be more effective when applied.

Public health significance: The development of a reliable method to identify and count germline mutations in radiation workers could be generalized to exposures from any form of environmental mutagen (e.g., chemicals). Such a generalized marker could be used to measure the effects of various toxic environmental exposures on specific individuals and to predict genetically determined illness conditions.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

I would like to express my profound thanks and appreciation to the members of my thesis committee. The guidance of Dr. Richard Day, Dr. Stephen Grant, Dr. Chien-Cheng (George) Tseng, was crucial in the development and structure of my thesis. Dr. Richard Day, my thesis advisor, deserves my special thanks and acknowledgements for it was because of his unwavering care and support that my thesis project came to fruition. Dr. Day served as my primary guide from the dawning of my project; encouraging me to write, strive for excellence in my analyses, and to remain steadfast in my goals. I would also like to thank Dr. Stephen for embracing the position of my academic advisor, and for taking the initiative to assist me in identifying a project for my thesis. To Dr. Tseng, I express my sincere thanks for his guidance and input in the statistical analyses selected for my thesis.

I would like to thank Joanne Pegher and Paige Shoemaker for their patience and assistance through the process of submitting of my thesis. Lastly, but most importantly, I would be remiss if I did not mention the care, and support of my loving family.

INTRODUCTION

Single nucleotide polymorphism (SNPs) genotyping arrays are a type of DNA microarray that is used to identify genetic variation in individuals and across populations [1, 2]. These SNP arrays are also used to profile somatic mutations in cancer, specifically loss of heterozygosity (LOH) events and amplifications and deletions of regions of DNA [2, 3, 4]. Genome SNP arrays assay oligonucleotides designed to blanket an entire genomic region of interest. Companies such as Illumina and Affymetrix have successfully designed SNP arrays at extremely high densities providing for hypothesis-free genome-wide scans for common polymorphisms associated with complex diseases [1, 2, 5, 6]. As one of these platforms, the Illumina Sentrix Human Hap 550 Genotyping BeadChip makes it possible to interrogate over 550,000 tagged SNPs located across the human genome and routinely provides genotype calls for upwards of 99.78% of all the SNP positions with advertised reproducibility rates on the order of 99.9%. The Illumina 550k BeadChip also routinely returns a small proportion (≈ 0.035%) of genotypes that demonstrate Mendelian inconsistencies (MIs) and/or null ("no call") values [7]. In the past, these MIs and null genotypes occurring in whole genome scan (WGS) have been ignored or excluded from analyses as probable technical errors [8]. Recently, however, a number of investigators [9, 10] have reported that MIs and null genotypes represent meaningful data that can be used to map the occurrence of copy number variants (CNVs) in the human genome. These investigators focused on clusters (or runs) of MIs that were ≥1Kb in size and which were described primarily as "deletion variants". Numerous, partially overlapping regions of CNV were identified by these investigators and validation studies indicated that 80-85% of these deletion variations could be confirmed using independent experimental methods [11].

Like Conrad et al. [9], we chose to focus on MIs occurring in WGS as a means of identifying potential deletion variants occurring in familial triads (two parents and a single offspring) by comparing populations of occupationally exposed and unexposed Russian radiation workers to Mormon subjects from the Foundation Jean Dausset (CEPH). Unlike other investigators who are interested in mapping common areas of CNV, our project concerns whether external environmental exposures (e.g., ionizing radiation) can affect the frequency of deletion variants (i.e., germline mutations) occurring in the offspring of highly exposed nuclear workers. This focus leads to a consideration of the possible origin of deletion variants identified in the worker offspring and possible methods for identifying different types of deletion variants.

MIs in WGS of familial triads using the HumanHap550 Genotyping BeadChip result from the fact that SNPs which are missing a single allele at a specific locus are designated ("called") by the Illumina BeadStation (IBS) software as homozygous for the present allele (e.g., A/- [actual]→AA [called]) [12]. Valid deletion variants occurring in the offspring of a triad may be of three mutually exclusive types:

1. Inherited parental deletion variants (PDVs): Established somatic CNVs, pre-existing in the parents in familial triads, may be directly passed down to the offspring. These established parental CNVs may represent spontaneous mutations occurring in the parental generation or previously established mutations handed down from the grand-parental generation or beyond. In the SNP data from the WGS, these inherited parental CNVs would appear as MIs in the triad offspring (e.g., A/- + BB →B/-; which appear in the IBS output as AA + BB →BB).
2. Spontaneous mutations (SMs): Spontaneous mutations represent de novo CNVs occurring for the first time in the triad's offspring generation. These would have the following form (AA + BB → B/-; IBS output: AA + BB → BB).
3. Germline mutations (GMs): Germline mutations represent de novo CNVs originating in a parental germ cell (i.e., sperm or ovum) that are passed down to the offspring generation. These CNVs may be spontaneous (i.e., SMs) or they may be induced by an external mutagens, such as

ionizing radiation. Since the mutation occurs in only a single parental germ cell, it will not appear in the parent's somatic cells or WGS. However, if the germ cell survives this mutation and a successful fertilization occurs, the CNV may be passed along to the offspring's generation and may appear in the offspring's WGS (AA + BB → B/-; IBS output: AA + BB →BB).

A reliable methodology for identifying the actual genotype from the observed genotype currently is not available, this paper attempt to test a probabilistic method adapting Bayesian theorem [13] by analysis of six putative clusters drawn from the SNP arrays for six Russian radiation worker family trios.

SUBJECTS AND METHODS

SUBJECTS

There are two subject groups. Twenty-four subjects who serve as the case group are Russian worker family trios (father, mother and one offspring). Among these 8 families there are 6 families where one of parents was exposed occupationally to ionizing radiation (three parents exposed to Gamma radiation with Mean ± SE = 1881.6 ± 479.8 (mGy) and to Alpha radiation with Alpha body burden Mean ± SE = 111.7 ± 52.5 (nCi); three parents exposed to Gamma radiation with Mean ± SE = 2539.5 ± 239.2 (mGy) and to Alpha radiation with Alpha body burden Mean ± SE = 17.8 ± 3.9 (nCi)), while two families lived in the city without a parental occupational exposure to ionizing radiation (radiation dose in the industry surrounding area was unknown but was subjected to be higher than those in living areas without any such industry around). None of the offspring of the 8 families had a history of employment in the ionizing radiation industry themselves. Other subjects who serve as the control group are 112 individuals who participated in the International HapMap Project, consisting of 16 CEPH family trios (n=48), 12 Yoruba family trios (n=36), 16 unrelated Japanese subjects, 12 unrelated Han Chinese subjects.

DNA SAMPLES AND PROCESS

The 24 DNA samples from the eight Russian trios were sent by express mail to the Core Genetics Laboratory (CGL) at the University of Pittsburgh (UPitt) where all of the samples were found to be in excellent condition and in amounts that made possible their assay with the Illumania Sentrix 550K BeadArray technology. 550K BeadArray assays were carried by the CGL at UPitt and the output was manipulated using Illumina's BeadStudio (IBS, ver.3). The 650K BeadArray data for 112 subjects from the International HapMap Project made available by the Illumina Corporation. These HapMap DNA samples were purchased from the Coriell Medical Research Institute and the assays were carried out in the Illumina laboratories. The output of the HapMap samples was also manipulated using the same IBS software.

FUNDAMENTALS OF GENOTYPING

The Illumina method queries genomic DNA with three probes per locus and creates DNA fragments that can be amplified by standard PCR methods using three universal primers. For each locus interrogated, the oligo mix contains two allele-specific and one locus-specific probe. The 3' end of the allele-specific probes are complementary to universal primer 1 and 2, the 5' end is complementary to the 3' end of the locus. Each allele-specific probe is complementary to the portion of genomic DNA that is 5 to 20 bases beyond the point of interrogation. This probe is terminated with the specific Illumacode that will be used to identify the locus, as well as the sequence for universal primer sequence 3. These probes are annealed to the genomic DNA, tag polymerase is added to close the gap between the allele-specific and locus-specific probes and the two fragments are ligated together. Probe fragments are separated from the genomic DNA and used to inoculate a PCR reaction. The primer mix for this PCR reaction consists of double stranded DNA of which one strand, containing the complementary strand is labeled with biotin. The biotinylated strand is removed and this single, fluorescently labeled strand hybridized to the array.

The BeadArray reader scans the hybridized array and determines the signal intensities for each dye at each bead location. Software mapping is then used to map the known location of each Illumacode on the Sentrix Array Matrix back to the locus being interrogated by that Illumacode and to match the dye intensities to the specific alleles [15].

IDENTIFICATION OF REGIONS OF CNV IN TERMS OF MIS

A review of the data showed that a significant proportion of these 24 subjects' MI's in the SNP arrays tended to cluster in specific areas of the genome. A "cluster" is defined in probabilistic manner. Given approximately 200 MIs per subject (Mean ± SE = 191 ± 14.6) and a total of 550, 000 marker positions, if MIs were random occurrences, it would be expected to see one MI about every 2,750 marker positions. Hence, the probability that two successive markers both are MI's is approximately 0.00036 and the probability that two MIs fall within 10 markers of each other is roughly 0.0036. Thus it was decided to try defining clusters using two criteria. Two rules were used to define a cluster: first, the source of the putative missing allele must be same (i.e., mother or father) through out the whole cluster; and, second, the cluster cannot contain any heterozygotes (i.e., AB genotype) in the offspring data. The latter criteria evolved because two adjacent MI's may be flanked by one or more homozygous markers (e.g. AA + AA → AA) for which allele loss cannot be determined using genotypes alone. A heterozygous genotype in offspring, however, marks a location where the IBS software detects an allele at each possible position and, thus, can be taken to represent the most extreme possible length of cluster, need to rationalize "cluster" with "deletion".

STATISTICAL PROCEDURES

In order to reasonably assess the probability of a missing allele at a specific location for a given cluster in the case group, we must compared the observed allele channel intensity to the characteristic mean level found at that location in a control group of 112 HapMap d subjects . First, the Call Score parameter of the IBS software was set to zero in order to maximize number of genotype calls. Once the regional CNVs were determined in terms of the defined putative clusters of MIs, six clusters with one each from each trio of six Russian families were randomly selected for the investigation. The output of the 112 individual HapMap Project data from the IBS software was sent to the SPSS software (SPSS, ver. 15), where mean and standard deviation of A allele and B allele channel intensity (or called x intensity, y intensity, respectively) for the each possible genotype AA, AB and BB at each locus are computed.

Due to the genotyping conventions of the IBS software mentioned earlier, an observed genotype AA is likely a genotype A/- and an observed genotype BB is probably a genotype B/-; however, it is believed that the proportions of the hemizgous (e.g., genotype A/-, or B/-) for a normal population of SNP data is very low across the whole SNP array, i.e., 0.035% of genotypes demonstrates MIs on 120 DNA samples from HapMap Project data [7]. Thus, the software is still considered to be an effective means of genome mapping especially for our control data. In order to attain our study goals it is necessary to figure out the actual genotype for the case data when the IBS output presents a genotype AA or BB. Besides a genotype AA, which might present an actual genotype AA or A/-, and a genotype BB, which might present an actual genotype BB or B/-, it is also necessary to determine whether the observed genotype AA or BB was possibly a genotype AB for the given locus. Thus, it is our prior belief that an observed genotype AA at a given locus for a case data would be possible a genotype AA, A/- or AB, while an observed genotype BB at a given locus would be possible a genotype BB, B/- or AB. In order to identify an actual genotype for the given locus, the Bayesian theorem is adapted to solve the encountered issues.

The Bayesian theorem [13] is simply a statement of conditional probability. Suppose that A1, …, Ak is any set of mutually exclusive and exhaustive events, and that events B and Ai are of special interest. Bayesian theorem for events provides a way to find the conditional probability of Ai given B in terms of

the conditional probability of B given Ai. For this reason, Bayesian theorem is also called a theorem about "inverse probability". Bayesian theorem for events is given by:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)},$$

for P(B) ≠ 0, where P(Ai) is a prior probability of event Ai, , in that it is one's degree of belief about event Ai prior to one's having any information about event B, while P(Ai|B) denotes a posterior probability of event Ai. The posterior probability represents our degree of belief about event Ai posterior to having information about B. P(B|Ai) and is the conditional probability of B given Ai and ∑jP(B|Aj)P(Aj) is the proposition that if { Aj : j = 1, 2, 3, ... } is a finite or countably infinite partition of a probability space.

In our study, in order to compute posterior probability of AA, A/- or AB for an observed AA and posterior probability of BB, B/- or AB for an observed BB, we need to have prior information of P(Ai) and P(B|Ai). The P (Ai) for an observed AA is given by P(AA), P(A/-) and P(AB) while the P(Ai) for an observed BB is given by P(BB), P(B/-) and P(AB). Since the proportions of three genotypes of AA, A/- and AB for an observed AA at any given locus are unknown as well as those for an observed BB, it is considered that the prior distribution of Ai is an uniform distribution, that is P(Ai) = 1 where 0<Ai<1; e.g., P(AA) =1, P(A/-)=1 and P(AB)=1 for an observed AA. Thus, posterior probability would depend upon the data only through P(B|Ai).

It is assumed that both A allele channel intensity and B allele channel intensity for any genotype at any given locus are a normal distribution, thus P(B|Ai) (P(obs|AA), P(obs|A/-), P(obs|AB) for an observed AA; P(obs|BB), P(obs|B/-), P(obs|AB) for an observed BB) is determined by one-sided p value obtained by computing z score. The z score formula is as the following:

$$Z = \frac{X - \mu}{\sigma}$$

where X is observed value, μ is mean value, and σ is standard deviation. For an observed AA genotype at a given locus, P(obs|AA) is the p value obtained by compared its x intensity to mean x intensity of AA at that locus, P(obs|A/-) is the p value obtained by compared its x intensity to mean x intensity of AB at that locus, and P(obs|AB) is the p value obtained by compared its y intensity to mean y intensity of AB at that locus. Similarly, for an observed BB genotype at a specific locus, P(obs|BB) is the p value obtained by compared its y intensity to mean y intensity of BB at the locus, P(obs|B/-) is the p value obtained by compared its y intensity to mean y intensity of AB at the locus, and P(obs|AB) is the p value obtained by compared its x intensity to mean x intensity of AB at the locus.

Other assumptions for the posterior probability calculation are:

- All observed genotypes at each locus in the SNPs array for the control data are "true" ;
- All observed genotype AB in the SNPs array for the case data are "true" as well;
- X intensity distribution of genotype A/- at a particular locus for the case data is the same as x intensity distribution of genotype AB at the same locus for the control data;
- Y intensity distribution of genotype B/- at a particular locus for the case data is the same as y intensity distribution of genotype AB at the same locus for the control data;

Once the posterior probabilities are computed, the greatest value of posterior probability among three determines the prior genotype at the given locus according to the Bayesian rule [6].

PDVS AND SMS/GMS DETERMINATIONS

Once a prior genotype at a specific locus is determined based on the posterior probability calculation, possible genotypes for offspring based on the laws of heredity can be determined, thus a missing allele

or gaining one more allele at the specific locus for offspring could be identified as inherited parental deletion variants (PDVs) or spontaneous mutations (SMs) / germline mutations (GMs). For example, given a specific locus, a paternal genotype is AA and a maternal genotype is B/-, thus possible genotypes for offspring are AB and A/- in a ratio of 1:1. If the offspring's prior genotype is B-, indicate that the missing allele is due to SMs/GMs; however, if the offspring's prior genotype is A/-, indicate that the missing allele is due to PDVs;

All selected data and the results of computation are stored in Excel spread sheets (Microsoft Excel, ver. 2003). Z score and its p value, and posterior probability are calculated by using STATA software (STATA, ver. 8).

RESULTS

SELECTED CLUSTERS

Six putative clusters with observed genotypic Mendelian inconsistencies for offspring for six Russian family trios were randomly selected. Information for each locus in the six clusters, including variables genotype (GType), A allele channel intensity (X), B allele channel intensity (Y) and so on, is presented in Table 1. Each variable displayed in Table 1 is described in detail in Table 2. As can be seen from Table 1, four clusters selected are due to a missing paternal allele while two out of six clusters are due to a missing maternal allele. It is noted that there is one locus with observed undetermined allele loss for triad number 217 and triad number 1966, respectively.

PRIOR INFORMATION

Tables 3.1.1 and 3.1.2 through Tables 8.1.1 and 8.1.2. present the prior information for the six clusters, respectively. For each table mentioned, the left section displays individual information for a family trio, the middle and right section display sample size, mean, standard deviation for each genotype at each given locus, as well as the computed z score and p value. Prior conditional probabilities calculated for each locus for the six clusters from the six trios are presented by p1, p2, p3 and p4. For an observed AA

genotype, p1 is P(obs|A/-) , p2 is P(obs|AB), and p3 is P(obs|AA); while for an observed BB genotype, p1 is P(obs|AB), p2 is P(obs|B/-), and p4 is P(obs|BB).

As can be seen from the tables mentioned above, sample sizes for the three genotypes AA, AB, and BB at the given loci are highly variable in the control population. In triad number 217, for example, the sample sizes for AB genotype at the given loci range from 13 to 57, sample sizes for the AA genotype range from 3 to 77, and sample sizes for the BB genotype range from 4 to 96.

POSTERIOR PROBABILITY AND PRIOR GENOTYPE

Tables 3.2 through 8.2 present the observed genotype in the left section, posterior probabilities of the three genotypes in the second section, a prior genotype in the third section, and the allele loss type in the last section for each family trio, respectively. Given the second section, for an observed genotype of AA or BB, p(I|obs) is the posterior probability of AB, p(II|obs) is the posterior probability of A/- or B/-, and p(III|obs) is the posterior probability of AA or BB. As can also be seen at the second section from the tables mentioned, a posterior probability of AB (p(I|obs)) for the all six clusters ranges from 0 to 0.01246, indicating that for an observed genotype of AA or BB, one can be almost certain that the genotype is not truly an AB.

Table 9 summarizes posterior probabilities for observed genotypes AAs and BBs in the six clusters. For observed AAs, there are about half of AA and half of A/-, while for observed BBs, two third are B/- and one third are BB. Overall, greater than 50% of the prior genotypes are determined by posterior probability values greater than 0.9.

MISSING ALLELE DUE TO PDVS OR SMS/GMS

Table 10 summarizes the frequencies of missing alleles for offspring of the six trio clusters due to PDVs or SMs/GMs. Among 26 loci with 24 observed allele losses all could be identified as either PDVs or SMs/GMs. Additionally, one observed undetermined allele loss in triad number 1966 is also identified as SMs/GMs. Unlike other clusters with mixed allele loss types of PDVs and SMs/GMs, all observed missing alleles in the cluster of trial number 521 and 832 are perfectly identified as SMs/GMs. Among the later two clusters, it is also noted that all three prior genotypes of offspring in the cluster of triad number 521 are anomalous (see Table 5.2). Given the genome index number 45511 for example, where paternal prior genotype is AB and maternal prior genotype is A/-, while offspring's prior genotype is BB. According to the rule of hereditary, possible genotype for offspring are AA, AB, A/- or B/-, so the one more gaining of B allele neither from the paternal nor maternal is of anomaly.

DISCUSSION

The Bayesian paradigm provides a consistent and understandable mechanism for solving inferential problems. Bayesian inferences are made in the form of probability statements about the unknown parameter [20], and are intuitively appealing and useful for solving the given genotyping problems. As can be seen from the results, we can identify the genotype in terms of allele loss at any specific locus using posterior probabilities and the laws of heredity. One can easily figure out genomic variation attributed to PDVs or SMs/GMs when studying the SNPs genotyping array for a family trio. Table 10 presents important result because 25/26 (96.15%) genotypes under investigation could be identified as either PDVs or SMs/GMs for offspring, and, in fact, 100% of the observed missing alleles could be indentified by combining probabilistic approach with the hereditary rule. Even one observed undetermined alleles (in the cluster of triad number 1966) could be identified by using the above methodology. And even more attractively, two clusters out of six are perfectly identified as SMs/GMs. For those clusters with mixed PDVs and SMs/GMs, one can distinguish whether missing alleles are carried from the parent(s) or newly occur due to somatic mutation or germline cell mutations in a given cluster. The ability to of distinguish PDVs and SMs/GMs makes it possible for further investigation of the causal relationship between external environmental exposures (e.g., ionizing radiation) and frequency of deletion variants occurring in the offspring of highly exposed nuclear workers (i.e., germline mutations).

While Bayesian approach provides a practical access to the problematic genotyping in the SNPs array in terms of the clusters of MIs, some deficiencies of the study must be noted.

First, the size of the training sample should be large enough to represent the true distribution of each SNP and to ensure accurate estimation [6]. The control sample from the International HapMap Project, consisting of only 112 DNA samples data, may not be large enough to provide stable estimates for all the specific genotypes, which can be seen from Tables 3.1.1 and 3.1.2 through Tables 8.1.1 and 8.1.2. Obviously, such small sample sizes across many loci may be insufficient to represent the true distribution at the given loci. Furthermore, it is concerned that three prior genotypes of offspring in the cluster of the triad number 521 are totally anomalous (see Table 5.2) as mentioned in the results aspect. This phenomenon could be due to real copy number variation or due to the procedure, e.g., very small sample sizes for certain genotypes at the given loci (see Table 5.1.1 and Table 5.1.2).

Second, the case group, which consists of Russian workers and their families, ethnically differs from the control group which consists of individuals of Yoruba, Han Chinese, Japanese and Mormons, whether this difference would affect the SNPs of two populations is unknown. In addition, the case group DNA samples were run in the PITT Core Labs while the control group DNA samples were carried out in the Illumina laboratories, this difference might carry in different systematic errors, if any, when the data is collected. Such different background regarding the ethnic and DNA sample process might possibly introduce bias when the data analysis is conducted.

Third, in the study an assumption of true genotype AA or BB in the control group is made. In fact, it is certain that in the control group some hemizygous SNPs, i.e. A/- or B/- are presented by IBS output genotype AA or BB [12] although its proportions are considered to be low [7], which could also probably introduce bias to the data analyses.

CONCLUSION

In summary, HumanHap550 Genotyping BeadChip provides a platform allowing for genotyping of SNPs at greater than 550,000 loci. Using this technology this study focuses on analysis of six putative clusters drawn from the six Russian radiation worker family trios. In order to recognize various deletion variants such as PDVs, SMs, and GMs, a methodology based on the Bayesian rule of identifying genotypes at given loci was tested. Given a posterior probability calculated, a prior genotype at a given locus is determined, along with the rules of heredity, allele loss in the offspring can be identified as PDVs or SMs/GMs. This allows for further investigation of causal relationship between external mutagen substances exposures and frequency of deletion variants occurring in the offspring of these highly exposed employees. The derived statistical method provides a practical method to deal with the problematic genotyping for the given cases although some weaknesses of this study are observed. Further statistical investigation and research, such as selecting appropriate training sample, increasing sample size, and cleaning the potential hemizygous SNPs (i.e., MIs) are urged in order to enhance the reliability and valid of the study.

APPENDIX: TABLES

Table 1 Original information of six putative clusters with observed genotype mismatches for six triads

Triad#	Clus	Chr	Position	Name	GenoIndex	MA	Gtype_1	X1	Y1	Gtype_2	X2	Y2	Gtype_3	X3	Y3
217	8	11	124580785	rs601414	372973	P	AA	1.6021	0.078	AB	1.3474	1.0525	BB	0.0234	1.27
217	8	11	124584492	rs11219971	372974	P	AA	1.2124	0.0275	AB	0.8929	0.7406	BB	0.005	0.9456
217	8	11	124587334	rs12808899	372975	P	BB	0	0.7272	AB	0.4903	0.6614	AA	0.638	0.0116
217	8	11	124588326	rs647737	372976	U	AA	0.2934	0.0152	AA	0.5228	0.0248	AA	0.3045	0.0149
217	8	11	124589952	rs4935911	372977	P	AA	0.6546	0.0028	AB	0.6254	0.7633	BB	0.0052	0.8339
457	2	2	208064035	rs918843	80139	P	AA	0.245	0	BB	0.0384	0.438	BB	0.0121	0.5143
457	2	2	208064167	rs918842	80140	P	AA	1.0636	0.0471	BB	0.0135	1.1834	BB	0.0096	1.1439
457	2	2	208064454	rs2551649	80141	P	BB	0.0085	0.7572	AA	0.6162	0.0197	AA	0.5982	0.0229
457	2	2	208065237	rs6755425	80142	P	BB	0.0212	0.8198	AA	0.469	0.0533	AA	0.4328	0.0624
457	2	2	208066083	rs959668	80143	P	AA	0.7063	0.0198	BB	0.0023	0.8493	BB	0	0.7901
521	3	2	14622841	rs2675899	45511	M	AB	0.2232	0.4991	AA	0.1731	0	BB	0.0125	0.5726
521	3	2	14625996	rs6734939	45512	M	AB	0.5255	0.7022	AA	0.6776	0	BB	0.0174	0.8606
521	3	2	14626938	rs10929916	45513	M	AB	0.8587	0.9278	AA	0.7158	0.0129	BB	0.0204	1.0302
832	1	1	173064490	rs2156853	27276	P	AA	0.4041	0	BB	0	0.9778	BB	0	0.7021
832	1	1	173067495	rs1540207	27277	P	AA	0.1096	0	BB	0	0.4055	BB	0	0.2615
832	1	1	173068262	rs2156854	27278	P	BB	0.0289	0.9657	AA	1.3032	0.1586	AA	0.8173	0.0794
905	5	3	163700688	rs6441503	118641	P	AA	0.2645	0	AB	0.2255	0.3329	BB	0	0.3244
905	5	3	163701543	rs1382203	118642	P	BB	0	0.4647	AB	0.2427	0.4459	AA	0.291	0.0025
905	5	3	163708811	rs7615219	118643	P	AA	0.3287	0	AB	0.2532	0.3691	BB	0	0.4525
905	5	3	163709228	rs6795447	118644	P	BB	0	0.2602	AB	0.1563	0.2599	AA	0.1736	0
905	5	3	163710564	rs10804797	118645	P	AA	0.1186	0	AB	0.1068	0.2461	BB	0.0017	0.2627
905	5	3	163712278	rs7627193	118646	P	AA	0.1015	0	AB	0.0811	0.1509	BB	0.0073	0.1862
1966	2	2	208064035	rs918843	80139	M	AB	0.1576	0.3683	BB	0.0302	0.4054	AA	0.1695	0.006
1966	2	2	208064167	rs918842	80140	U	AA	1.4796	0.0812	AA	1.1261	0.0633	AA	1.0069	0.054
1966	2	2	208064454	rs2551649	80141	M	AB	0.5057	0.7003	AA	0.5934	0.0096	BB	0	0.7988
1966	2	2	208065237	rs6755425	80142	M	AB	0.4058	0.7364	AA	0.491	0.0576	BB	0.0197	0.7953

Table 2 Description of variables presented in Table 1

Column	Variable Name	Description of variables	Note
1	Triad#	Triad number	
2	Clus	Putative cluster	
3	Chr	Chromosome number	
4	Position	Genome position	
5	Name	Genome name	
6	Genoindex	Genome index	
7	MA	Parental missing allele	value "P"=paternal; value "M"=maternal; value "U"= undetermined
8	Gtype_1	Paternal genotype	
9	X1	Paternal A allele channel intensity	
10	Y1	Paternal B allele channel intensity	
11	Gtype_2	Maternal genotype	
12	X2	Maternal A allele channel intensity	
13	Y2	Maternal B allele channel intensity	
14	Gtype_3	Offspring genotype	
15	X3	Offspring A allele channel intensity	
16	Y3	Offspring B allele channel intensity	

Table 3.1.1 Prior information for genotype of each locus for the triad number 217 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
217	FA	372973		AA	1.6021	0.078	57	1.701	0.155	-0.63806	0.261716	57	1.247	0.074	-15.7973	1.62E-56
217	FA	372974		AA	1.2124	0.0275	31	0.958	0.127	2.003149	0.022581	31	0.805	0.086	-9.0407	7.78E-20
217	FA	372975		BB	0	0.7272	33	0.371	0.041	-9.04878	7.23E-20	33	0.662	0.042	1.55238	0.060286
217	FA	372976		AA	0.2934	0.0152	53	0.226	0.029	2.324138	0.010059	53	0.257	0.023	-10.513	3.76E-26
217	FA	372977		AA	0.6546	0.0028	13	0.431	0.133	1.681203	0.046362	13	0.665	0.204	-3.24608	0.000585
217	MO	372973		AB	1.3474	1.0525	57	1.701	0.155	-2.28129	0.011266	57	1.247	0.074	-2.62838	0.00429
217	MO	372974		AB	0.8929	0.7406	31	0.958	0.127	-0.5126	0.304116	31	0.805	0.086	-0.74884	0.226978
217	MO	372975		AB	0.4903	0.6614	33	0.371	0.041	2.909756	0.001809	33	0.662	0.042	-0.01429	0.494301
217	MO	372976		AA	0.5228	0.0248	53	0.226	0.029	10.23448	6.95E-25	53	0.257	0.023	-10.0957	2.89E-24
217	MO	372977		AB	0.6254	0.7633	13	0.431	0.133	1.461654	0.071918	13	0.665	0.204	0.481863	0.314952
217	OFF	372973	P	BB	0.0234	1.27	57	1.701	0.155	-10.8232	1.34E-27	57	1.247	0.074	0.310811	0.377972
217	OFF	372974	P	BB	0.005	0.9456	31	0.958	0.127	-7.50394	3.10E-14	31	0.805	0.086	1.634883	0.051037
217	OFF	372975	P	AA	0.638	0.0116	33	0.371	0.041	6.512196	3.70E-11	33	0.662	0.042	-15.4857	2.17E-54
217	OFF	372976	U	AA	0.3045	0.0149	53	0.226	0.029	2.706897	0.003396	53	0.257	0.023	-10.5261	3.28E-26
217	OFF	372977	P	BB	0.0052	0.8339	13	0.431	0.133	-3.2015	0.000684	13	0.665	0.204	0.827941	0.203852

Note: variable Rela=relationship among family; "FA"=father; "MO"=mother; "OFF"=offspring.

Table 3.1.2 Prior information for genotype of each locus for the triad number 217 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
217	FA	372973		AA	1.6021	0.078	24	2.779	0.268	-4.39142	5.63E-06	31	1.72	0.15		
217	FA	372974		AA	1.2124	0.0275	77	1.869	0.212	-3.09717	0.000977	4	1.215	0.095		
217	FA	372975		BB	0	0.7272	72	0.812	0.089			7	1.039	0.051	6.11373	4.87E-10
217	FA	372976		AA	0.2934	0.0152	47	0.437	0.049	-2.93061	0.001691	12	0.445	0.039		
217	FA	372977		AA	0.6546	0.0028	3	0.98	0.041	-7.93659	1.04E-15	96	1.041	0.142		
217	MO	372973		AB	1.3474	1.0525	24	2.779	0.268			31	1.72	0.15		
217	MO	372974		AB	0.8929	0.7406	77	1.869	0.212			4	1.215	0.095		
217	MO	372975		AB	0.4903	0.6614	72	0.812	0.089			7	1.039	0.051		
217	MO	372976		AA	0.5228	0.0248	47	0.437	0.049	1.751021	0.039971	12	0.445	0.039		
217	MO	372977		AB	0.6254	0.7633	3	0.98	0.041			96	1.041	0.142		
217	OFF	372973	P	BB	0.0234	1.27	24	2.779	0.268			31	1.72	0.15	-3	0.00135
217	OFF	372974	P	BB	0.005	0.9456	77	1.869	0.212			4	1.215	0.095	2.83579	0.002286
217	OFF	372975	P	AA	0.638	0.0116	72	0.812	0.089	-1.95506	0.025288	7	1.039	0.051		
217	OFF	372976	U	AA	0.3045	0.0149	47	0.437	0.049	-2.70408	0.003425	12	0.445	0.039		
217	OFF	372977	P	BB	0.0052	0.8339	3	0.98	0.041			96	1.041	0.142	1.45845	0.072358

Table 3.2 Identification of genotype and allele loss for each locus for the triad number 217 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
217	FA	372973		AA	0	1	0	A/-		
217	FA	372974		AA	0	0.95853	0.04147	A/-		
217	FA	372975		BB	0	1	0	B/-		
217	FA	372976		AA	0	0.85609	0.14391	A/-		
217	FA	372977		AA	0.01246	0.98754	0	A/-		
217	MO	372973		AB				AB		
217	MO	372974		AB				AB		
217	MO	372975		AB				AB		
217	MO	372976		AA	0	0	1	AA		
217	MO	372977		AB				AB		
217	OFF	372973	P	BB	0	0.99644	0.00356	B/-	AA,AB,A/-,B/-	PDVs
217	OFF	372974	P	BB	0	0.95713	0.04287	B/-	AA,AB,A/-,B/-	PDVs
217	OFF	372975	P	AA	0	0	1	AA	AB,BB,A/-,B/-	SMs/GMs
217	OFF	372976	U	AA	0	0.49787	0.50213	AA	2AA,2A/-	undetermined
217	OFF	372977	P	BB	0.00247	0.73621	0.26132	B/-	AA,AB,A/-,B/-	PDVs

Table 4.1.1 Prior information for genotype of each locus for the triad number 457 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
457	FA	80139		AA	0.245	0	36	0.183	0.023	2.695652	0.003513	36	0.395	0.03	-13.1667	6.82E-40
457	FA	80140		AA	1.0636	0.0471	26	0.779	0.234	1.216239	0.111947	26	0.894	0.262	-3.23244	0.000614
457	FA	80141		BB	0.0085	0.7572	36	0.572	0.053	-10.6321	1.06E-26	36	0.765	0.043	-0.18139	0.428029
457	FA	80142		BB	0.0212	0.8198	37	0.414	0.038	-10.3368	2.40E-25	37	0.764	0.046	1.213044	0.112557
457	FA	80143		AA	0.7063	0.0198	28	0.666	0.218	0.184862	0.426668	28	0.714	0.204	-3.40294	0.000333
457	MO	80139		BB	0.0384	0.438	36	0.183	0.023	-6.28696	1.62E-10	36	0.395	0.03	1.433333	0.075881
457	MO	80140		BB	0.0135	1.1834	26	0.779	0.234	-3.27137	0.000535	26	0.894	0.262	1.10458	0.134671
457	MO	80141		AA	0.6162	0.0197	36	0.572	0.053	0.833961	0.202151	36	0.765	0.043	-17.3326	1.34E-67
457	MO	80142		AA	0.469	0.0533	37	0.414	0.038	1.447369	0.073897	37	0.764	0.046	-15.45	3.77E-54
457	MO	80143		BB	0.0023	0.8493	28	0.666	0.218	-3.0445	0.001165	28	0.714	0.204	0.663236	0.25359
457	OFF	80139	P	BB	0.0121	0.5143	36	0.183	0.023	-7.43044	5.41E-14	36	0.395	0.03	3.976666	3.49E-05
457	OFF	80140	P	BB	0.0096	1.1439	26	0.779	0.234	-3.28803	0.000504	26	0.894	0.262	0.953817	0.170088
457	OFF	80141	P	AA	0.5982	0.0229	36	0.572	0.053	0.49434	0.310533	36	0.765	0.043	-17.2581	4.86E-67
457	OFF	80142	P	AA	0.4328	0.0624	37	0.414	0.038	0.494737	0.310393	37	0.764	0.046	-15.2522	7.96E-53
457	OFF	80143	P	BB	0	0.7901	28	0.666	0.218	-3.05505	0.001125	28	0.714	0.204	0.373039	0.35456

Table 4.1.2 Prior information for genotype of each locus for the triad number 457 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
457	FA	80139		AA	0.245	0	57	0.311	0.099	0.66667	0.252493	19	0.574	0.115		
457	FA	80140		AA	1.0636	0.0471	81	1.473	0.359	1.14039	0.127062	5	1.256	0.179		
457	FA	80141		BB	0.0085	0.7572	22	0.843	0.428			54	1.059	0.145	2.08138	0.0187
457	FA	80142		BB	0.0212	0.8198	20	0.565	0.326			55	1.016	0.157	1.24968	0.105708
457	FA	80143		AA	0.7063	0.0198	8	0.943	0.497	0.47626	0.316945	76	1.091	0.153		
457	MO	80139		BB	0.0384	0.438	57	0.311	0.099			19	0.574	0.115	1.18261	0.118482
457	MO	80140		BB	0.0135	1.1834	81	1.473	0.359			5	1.256	0.179	0.40559	0.342523
457	MO	80141		AA	0.6162	0.0197	22	0.843	0.428	0.52991	0.298088	54	1.059	0.145		
457	MO	80142		AA	0.469	0.0533	20	0.565	0.326	0.29448	0.384196	55	1.016	0.157		
457	MO	80143		BB	0.0023	0.8493	8	0.943	0.497			76	1.091	0.153	1.57974	0.057083
457	OFF	80139	P	BB	0.0121	0.5143	57	0.311	0.099			19	0.574	0.115	0.51913	0.301835
457	OFF	80140	P	BB	0.0096	1.1439	81	1.473	0.359			5	1.256	0.179	0.62626	0.265573
457	OFF	80141	P	AA	0.5982	0.0229	22	0.843	0.428	0.57196	0.283674	54	1.059	0.145		
457	OFF	80142	P	AA	0.4328	0.0624	20	0.565	0.326	0.40552	0.342547	55	1.016	0.157		
457	OFF	80143	P	BB	0	0.7901	8	0.943	0.497			76	1.091	0.153	1.96667	0.024611

Table 4.2 Identification of genotype and allele loss for each locus for the triad number 457 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
457	FA	80139		AA	0	0.01372	0.98628	AA		
457	FA	80140		AA	0.00256	0.46718	0.53026	AA		
457	FA	80141		BB	0	0.95814	0.04186	B/-		
457	FA	80142		BB	0	0.51569	0.48431	B/-		
457	FA	80143		AA	0.000448	0.57352	0.42603	A/-		
457	MO	80139		BB	0	0.39041	0.60959	BB		
457	MO	80140		BB	0.00112	0.2819	0.71698	BB		
457	MO	80141		AA	0	0.40411	0.59589	AA		
457	MO	80142		AA	0	0.16131	0.83869	AA		
457	MO	80143		BB	0.00374	0.81321	0.18305	B/-		
457	OFF	80139	P	BB	0	0.00012	0.99988	BB	4AB	SMs/GMs
457	OFF	80140	P	BB	0.00115	0.38997	0.60888	BB	4AB	SMs/GMs
457	OFF	80141	P	AA	0	0.5226	0.47739	A/-	2AB, 2A/-	PDVs
457	OFF	80142	P	AA	0	0.47537	0.52463	AA	2AB, 2A/-	SMs/GMs
457	OFF	80143	P	BB	0.00297	0.93232	0.06471	B/-	AB, A/-,B/-,--	PDVs

Table 5.1.1 Prior information for genotype of each locus for the triad number 521 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
521	FA	45511		AB	0.2232	0.4991	60	0.18	0.021	2.05714	0.019836	60	0.36	0.031	4.487096	3.61E-06
521	FA	45512		AB	0.5255	0.7022	24	0.47	0.039	1.42308	0.077357	24	0.727	0.039	-0.6359	0.262422
521	FA	45513		AB	0.8587	0.9278	49	0.679	0.053	3.39057	0.000349	49	0.739	0.046	4.104347	2.03E-05
521	MO	45511		AA	0.1731	0	60	0.18	0.021	0.32857	0.37124	60	0.36	0.031	-11.6129	1.77E-31
521	MO	45512		AA	0.6776	0	24	0.47	0.039	5.32308	5.10E-08	24	0.727	0.039	-18.641	7.47E-78
521	MO	45513		AA	0.7158	0.0129	49	0.679	0.053	0.69434	0.243735	49	0.739	0.046	-15.7848	1.98E-56
521	OFF	45511	M	BB	0.0125	0.5726	60	0.18	0.021	7.97619	7.55E-16	60	0.36	0.031	6.858064	3.49E-12
521	OFF	45512	M	BB	0.0174	0.8606	24	0.47	0.039	11.6051	1.94E-31	24	0.727	0.039	3.425641	0.000307
521	OFF	45513	M	BB	0.0204	1.0302	49	0.679	0.053	12.4264	9.39E-36	49	0.739	0.046	6.330434	1.22E-10

Table 5.1.2 Prior information for genotype for the triad number 521 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
521	FA	45511		AB	0.2232	0.4991	14	0.359	0.082	-1.6561		38	0.566	0.074	0.90405	
521	FA	45512		AB	0.5255	0.7022	4	0.906	0.238	1.59874		84	1.083	0.085	-4.48	
521	FA	45513		AB	0.8587	0.9278	31	1.315	0.2	-2.2815		32	1.044	0.073	1.59178	
521	MO	45511		AA	0.1731	0	14	0.359	0.082	2.26707	0.011693	38	0.566	0.074		
521	MO	45512		AA	0.6776	0	4	0.906	0.238	0.95966	0.168612	84	1.083	0.085		
521	MO	45513		AA	0.7158	0.0129	31	1.315	0.2	-2.996	0.001368	32	1.044	0.073		
521	OFF	45511	M	BB	0.0125	0.5726	14	0.359	0.082			38	0.566	0.074	0.08919	0.464466
521	OFF	45512	M	BB	0.0174	0.8606	4	0.906	0.238			84	1.083	0.085	2.61647	0.004442
521	OFF	45513	M	BB	0.0204	1.0302	31	1.315	0.2			32	1.044	0.073	0.18904	0.42503

Table 5.2 Identification of genotype and allele loss for each locus for the triad number 521 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
521	FA	45511		AB				AB		
521	FA	45512		AB				AB		
521	FA	45513		AB				AB		
521	MO	45511		AA	0	0.96946	0.03054	A/-		
521	MO	45512		AA	0	0	1	AA		
521	MO	45513		AA	0	0.99442	0.00558	A/-		
521	OFF	45511	M	BB	0	0	1	BB	AA, AB, A/-,B/-	SMs/GMs
521	OFF	45512	M	BB	0	0.06467	0.93533	BB	2AA,2AB	SMs/GMs
521	OFF	45513	M	BB	0	0	1	BB	AA, AB, A/-,B/-	SMs/GMs

Table 6.1.1 Prior information for genotype of each locus for the triad number 832 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
832	FA	27276		AA	0.4041	0	28	0.359	0.035	1.288571	0.098774	28	0.665	0.054	-12.3148	3.77E-35
832	FA	27277		AA	0.1096	0	28	0.106	0.012	0.3	0.382089	28	0.188	0.022	-8.54545	6.40E-18
832	FA	27278		BB	0.0289	0.9657	28	0.743	0.069	-10.3493	2.11E-25	28	1.012	0.062	-0.74677	0.2276
832	MO	27276		BB	0	0.9778	28	0.359	0.035	-10.2571	5.50E-25	28	0.665	0.054	5.792592	3.47E-09
832	MO	27277		BB	0	0.4055	28	0.106	0.012	-8.83333	5.08E-19	28	0.188	0.022	9.886364	2.39E-23
832	MO	27278		AA	1.3032	0.1586	28	0.743	0.069	8.118841	2.35E-16	28	1.012	0.062	-13.7645	2.08E-43
832	OFF	27276	P	BB	0	0.7021	28	0.359	0.035	-10.2571	5.50E-25	28	0.665	0.054	0.687036	0.24603
832	OFF	27277	P	BB	0	0.2615	28	0.106	0.012	-8.83333	5.08E-19	28	0.188	0.022	3.340909	0.000418
832	OFF	27278	P	AA	0.8173	0.0794	28	0.743	0.069	1.076812	0.140782	28	1.012	0.062	-15.0419	1.95E-51

Table 6.1.2 Prior information for genotype of each locus for the triad number 832 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
832	FA	27276		AA	0.4041	0	30	0.661	0.284	-0.90458	0.182845	53	0.902	0.156		
832	FA	27277		AA	0.1096	0	30	0.167	0.077	-0.74545	0.227999	54	0.281	0.073		
832	FA	27278		BB	0.0289	0.9657	57	1.17	0.439			27	1.316	0.139	-2.52014	0.005865
832	MO	27276		BB	0	0.9778	30	0.661	0.284			53	0.902	0.156	0.485897	0.31352
832	MO	27277		BB	0	0.4055	30	0.167	0.077			54	0.281	0.073	1.70548	0.044052
832	MO	27278		AA	1.3032	0.1586	57	1.17	0.439	0.303417	0.380786	27	1.316	0.139		
832	OFF	27276	P	BB	0	0.7021	30	0.661	0.284			53	0.902	0.156	-1.28141	0.100025
832	OFF	27277	P	BB	0	0.2615	30	0.167	0.077			54	0.281	0.073	-0.26712	0.394687
832	OFF	27278	P	AA	0.8173	0.0794	57	1.17	0.439	-0.80342	0.210867	27	1.316	0.139		

Table 6.2 Identification of genotype and allele loss for each locus for the triad number 832 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
832	FA	27276		AA	0	0.35073	0.64927	AA		
832	FA	27277		AA	0	0.62628	0.37372	A/-		
832	FA	27278		BB	0	0.97488	0.02512	B/-		
832	MO	27276		BB	0	0	1	BB		
832	MO	27277		BB	0	0	1	BB		
832	MO	27278		AA	0	0	1	AA		
832	OFF	27276	P	BB	0	0.71096	0.28904	B/-	4AB	SMs/GMs
832	OFF	27277	P	BB	0	0.00106	0.99894	BB	2AB, 2B/-	SMs/GMs
832	OFF	27278	P	AA	0	0.40034	0.59966	AA	2AB, 2A/-	SMs/GMs

Table 7.1.1 Prior information for genotype of each locus for the triad number 905 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
905	FA	118641		AA	0.2645	0	45	0.189	0.032	2.359375	0.009153	45	0.293	0.043	-6.81395	4.75E-12
905	FA	118642		BB	0	0.4647	45	0.236	0.032	-7.375	8.22E-14	45	0.449	0.054	0.290741	0.385625
905	FA	118643		AA	0.3287	0	29	0.288	0.057	0.714035	0.237603	29	0.436	0.066	-6.60606	1.97E-11
905	FA	118644		BB	0	0.2602	29	0.193	0.035	-5.51429	1.75E-08	29	0.311	0.036	-1.41111	0.079106
905	FA	118645		AA	0.1186	0	30	0.108	0.026	0.407692	0.34175	30	0.223	0.05	-4.46	4.10E-06
905	FA	118646		AA	0.1015	0	29	0.12	0.037	-0.5	0.308538	29	0.19	0.046	-4.13044	1.81E-05
905	MO	118641		AB	0.2255	0.3329	45	0.189	0.032	1.140625	0.127013	45	0.293	0.043	0.927906	0.176728
905	MO	118642		AB	0.2427	0.4459	45	0.236	0.032	0.209375	0.417078	45	0.449	0.054	-0.05741	0.47711
905	MO	118643		AB	0.2532	0.3691	29	0.288	0.057	-0.61053	0.270757	29	0.436	0.066	-1.01364	0.155378
905	MO	118644		AB	0.1563	0.2599	29	0.193	0.035	-1.04857	0.147188	29	0.311	0.036	-1.41944	0.077885
905	MO	118645		AB	0.1068	0.2461	30	0.108	0.026	-0.04615	0.481594	30	0.223	0.05	0.462	0.322041
905	MO	118646		AB	0.0811	0.1509	29	0.12	0.037	-1.05135	0.146549	29	0.19	0.046	-0.85	0.197663
905	OFF	118641	P	BB	0	0.3244	45	0.189	0.032	-5.90625	1.75E-09	45	0.293	0.043	0.730232	0.232624
905	OFF	118642	P	AA	0.291	0.0025	45	0.236	0.032	1.71875	0.04283	45	0.449	0.054	-8.26852	6.78E-17
905	OFF	118643	P	BB	0	0.4525	29	0.288	0.057	-5.05263	2.18E-07	29	0.436	0.066	0.25	0.401294
905	OFF	118644	P	AA	0.1736	0	29	0.193	0.035	-0.55429	0.289692	29	0.311	0.036	-8.63889	2.84E-18
905	OFF	118645	P	BB	0.0017	0.2627	30	0.108	0.026	-4.08846	2.17E-05	30	0.223	0.05	0.794	0.213598
905	OFF	118646	P	BB	0.0073	0.1862	29	0.12	0.037	-3.04595	0.00116	29	0.19	0.046	-0.08261	0.467081

Table 7.1.2 Prior information for genotype of each locus for the triad number 905 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
905	FA	118641		AA	0.2645	0	51	0.39	0.099	1.26768	0.102457	16	0.531	0.053		
905	FA	118642		BB	0	0.4647	17	0.478	0.131			50	0.702	0.089	2.66629	0.003835
905	FA	118643		AA	0.3287	0	77	0.582	0.135	-1.8763	0.030307	6	0.75	0.082		
905	FA	118644		BB	0	0.2602	7	0.356	0.173			76	0.511	0.079	3.17468	0.00075
905	FA	118645		AA	0.1186	0	76	0.227	0.046	2.35652	0.009223	6	0.405	0.019		
905	FA	118646		AA	0.1015	0	77	0.227	0.068	1.84559	0.032476	6	0.358	0.063		
905	MO	118641		AB	0.2255	0.3329	51	0.39	0.099			16	0.531	0.053		
905	MO	118642		AB	0.2427	0.4459	17	0.478	0.131			50	0.702	0.089		
905	MO	118643		AB	0.2532	0.3691	77	0.582	0.135			6	0.75	0.082		
905	MO	118644		AB	0.1563	0.2599	7	0.356	0.173			76	0.511	0.079		
905	MO	118645		AB	0.1068	0.2461	76	0.227	0.046			6	0.405	0.019		
905	MO	118646		AB	0.0811	0.1509	77	0.227	0.068			6	0.358	0.063		
905	OFF	118641	P	BB	0	0.3244	51	0.39	0.099			16	0.531	0.053	3.89811	4.85E-05
905	OFF	118642	P	AA	0.291	0.0025	17	0.478	0.131	1.42748	0.076721	50	0.702	0.089		
905	OFF	118643	P	BB	0	0.4525	77	0.582	0.135			6	0.75	0.082	3.62805	0.000143
905	OFF	118644	P	AA	0.1736	0	7	0.356	0.173	1.05434	0.145865	76	0.511	0.079		
905	OFF	118645	P	BB	0.0017	0.2627	76	0.227	0.046			6	0.405	0.019	7.48947	3.46E-14
905	OFF	118646	P	BB	0.0073	0.1862	77	0.227	0.068			6	0.358	0.063	2.72698	0.003196

Table 7.2 Identification of genotype and allele loss for each locus for the triad number 905 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
905	FA	118641		AA	0	0.082	0.918	AA		
905	FA	118642		BB	0	0.99015	0.00985	B/-		
905	FA	118643		AA	0	0.88688	0.11312	A/-		
905	FA	118644		BB	0	0.99061	0.00939	B/-		
905	FA	118645		AA	0	0.97372	0.02628	A/-		
905	FA	118646		AA	0.000053	0.90472	0.09523	A/-		
905	MO	118641		AB				AB		
905	MO	118642		AB				AB		
905	MO	118643		AB				AB		
905	MO	118644		AB				AB		
905	MO	118645		AB				AB		
905	MO	118646		AB				AB		
905	OFF	118641	P	BB	0	0.99979	0.000208	B/-	2AA,2AB	SMs/GMs
905	OFF	118642	P	AA	0	0.35826	0.64174	AA	AB,BB,A/-,B/-	SMs/GMs
905	OFF	118643	P	BB	0	0.99964	0.000356	B/-	AA,AB,A/-,B/-	PDVs
905	OFF	118644	P	AA	0	0.66511	0.33489	A/-	AB,BB,A/-,B/-	PDVs
905	OFF	118645	P	BB	0	1	0	B/-	AA,AB,A/-,B/-	PDVs
905	OFF	118646	P	BB	0.00246	0.99076	0.00678	B/-	AA,AB,A/-,B/-	PDVs

Table 8.1.1 Prior information for genotype of each locus for the triad number 1966 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n1	ab_x	ab_x_sd	z_ab_x	p1	n2	ab_y	ab_y_sd	z_ab_y	p2
1966	FA	80139		AB	0.1576	0.3683	36	0.183	0.023	-1.10435	0.134721	36	0.395	0.03	-0.89	0.186733
1966	FA	80140		AA	1.4796	0.0812	26	0.779	0.234	2.994017	0.001377	26	0.894	0.262	-3.10229	0.00096
1966	FA	80141		AB	0.5057	0.7003	36	0.572	0.053	-1.25094	0.105477	36	0.765	0.043	-1.50465	0.066207
1966	FA	80142		AB	0.4058	0.7364	37	0.414	0.038	-0.21579	0.414576	37	0.764	0.046	-0.6	0.274253
1966	MO	80139		BB	0.0302	0.4054	36	0.183	0.023	-6.64348	1.53E-11	36	0.395	0.03	0.346667	0.364421
1966	MO	80140		AA	1.1261	0.0633	26	0.779	0.234	1.483333	0.068993	26	0.894	0.262	-3.17061	0.000761
1966	MO	80141		AA	0.5934	0.0096	36	0.572	0.053	0.403773	0.34319	36	0.765	0.043	-17.5674	2.19E-69
1966	MO	80142		AA	0.491	0.0576	37	0.414	0.038	2.026316	0.021366	37	0.764	0.046	-15.3565	1.60E-53
1966	OFF	80139	M	AA	0.1695	0.006	36	0.183	0.023	-0.58696	0.278616	36	0.395	0.03	-12.9667	9.45E-39
1966	OFF	80140	U	AA	1.0069	0.054	26	0.779	0.234	0.973932	0.165045	26	0.894	0.262	-3.20611	0.000673
1966	OFF	80141	M	BB	0	0.7988	36	0.572	0.053	-10.7925	1.87E-27	36	0.765	0.043	0.786047	0.21592
1966	OFF	80142	M	BB	0.0197	0.7953	37	0.414	0.038	-10.3763	1.59E-25	37	0.764	0.046	0.680435	0.248115

Table 8.1.2 Prior information for genotype of each locus for the triad number 1966 cluster

Triad#	Rela	GenoIndex	MA	GType	X	Y	n3	aa_x	aa_x_sd	z_aa_x	p3	n4	bb_y	bb_y_sd	z_bb_y	p4
1966	FA	80139		AB	0.1576	0.3683	57	0.311	0.099	-1.5495		19	0.574	0.115	-1.7887	
1966	FA	80140		AA	1.4796	0.0812	81	1.473	0.359	0.018384	0.492666	5	1.256	0.179		
1966	FA	80141		AB	0.5057	0.7003	22	0.843	0.428	-0.78808		54	1.059	0.145	2.47379	
1966	FA	80142		AB	0.4058	0.7364	20	0.565	0.326	-0.48834		55	1.016	0.157	1.78089	
1966	MO	80139		BB	0.0302	0.4054	57	0.311	0.099			19	0.574	0.115	1.46609	0.071312
1966	MO	80140		AA	1.1261	0.0633	81	1.473	0.359	-0.9663	0.166948	5	1.256	0.179		
1966	MO	80141		AA	0.5934	0.0096	22	0.843	0.428	-0.58318	0.279887	54	1.059	0.145		
1966	MO	80142		AA	0.491	0.0576	20	0.565	0.326	-0.22699	0.410214	55	1.016	0.157		
1966	OFF	80139	M	AA	0.1695	0.006	57	0.311	0.099	-1.42929	0.07646	19	0.574	0.115		
1966	OFF	80140	U	AA	1.0069	0.054	81	1.473	0.359	-1.29833	0.097087	5	1.256	0.179		
1966	OFF	80141	M	BB	0	0.7988	22	0.843	0.428			54	1.059	0.145	1.79448	0.036368
1966	OFF	80142	M	BB	0.0197	0.7953	20	0.565	0.326			55	1.016	0.157	1.40573	0.079902

Table 8.2 Identification of genotype and allele loss for each locus for the triad number 1966 cluster

Triad#	Rela	GenoIndex	MA	GType	p(I|obs)	p(II|obs)	p(III|obs)	prior gtype	possible gtypes for offspring	allele loss
1966	FA	80139		AB				AB		
1966	FA	80140		AA	0.00194	0.00278	0.99528	AA		
1966	FA	80141		AB				AB		
1966	FA	80142		AB				AB		
1966	MO	80139		BB	0	0.83634	0.16366	B/-		
1966	MO	80140		AA	0.00322	0.29148	0.70531	AA		
1966	MO	80141		AA	0	0.5508	0.4492	A/-		
1966	MO	80142		AA	0	0.04951	0.95049	AA		
1966	OFF	80139	M	AA	0	0.78467	0.21533	A/-	AB,BB,A/-,B/-	PDVs
1966	OFF	80140	U	AA	0.00256	0.62801	0.36943	A/-	4AA	SMs/GMs
1966	OFF	80141	M	BB	0	0.85585	0.14415	B/-	AA,AB,A/-,B/-	PDVs
1966	OFF	80142	M	BB	0	0.75641	0.24359	B/-	2AA,2AB	SMs/GMs

Table 9 Frequencies of posterior probability (Pr) for determination of actual genotypes for observed AAs or BBs within the six trio clusters

Observed GType	Prior GType	Pr ≥ 0.50	Pr ≥ 0.60	Pr ≥ 0.70	Pr ≥ 0.80	Pr ≥ 0 .90	Pr =1.00	Total (%)
AA	A/-	3	3	1	2	6	1	16 (48.48)
	AA	5	2	1	1	4	4	17 (51.52)
BB	B/-	1	0	3	3	10	2	19 (65.52)
	BB	0	2	1	0	3	4	10 (34.48)
Total (%)		9 (14.52)	7 (11.29)	6 (9.68)	6 (9.68)	23 (37.10)	11 (17.74)	62 (100.00)

Table 10 Missing alleles due to PDVs or SMs/GMs for offspring

Triad #	Observed missing allele (%)	PDVs (%)	SMs/GMs (%)	Total (%)
217	4/5 (80.00)	3/5 (60.00)	1/5 (20.00)	4/5 (80.00)
457	5/5 (100.00)	1/5 (20.00)	4/5 (80.00)	5/5 (100.00)
521	3/3 (100.00)	0/3 (0)	3/3 (100.00)	3/3 (100.00)
832	3/3 (100.00)	0/3 (0)	3/3 (100.00)	3/3 (100.00)
905	6/6 (100.00)	4/6 (66.67)	2/6 (33.33)	6/6 (100.00)
1966	3/4 (75.00)	2/4 (50.00)	2/4 (50.00)	4/4 (100.00)
Total %)	24/26 (92.31)	10/26 (38.46)	15/26 (57.69)	25/26 (96.15)

BIBLIOGRAPHY

1. Wang L, Luhm R, Lei M. (2007). SNP and mutation analysis. **Adv Exp Med Biol**. 593:105-16.

2. Craig DW & Stephan DA. (2005). Applications of whole-genome high-density SNP genotyping. **Expert Rev Mol Diagn.** 5(2):159-70.

3. Zheng HT, Peng ZH, Li S, & He L. (2005). Loss of heterozygosity analyzed by single nucleotide polymorphism array in cancer. **World J Gastroenterol** 11(43):6740-4.

4. Dutt A. & Beroukhim R. (2007). Single nucleotide polymorphism array analysis of cancer. **Curr Opin Oncol.** 19(1):43-9.

5. Infinium Genotyping Data Analysis http://www.illumina.com/downloads/GTDataAnalysis_TechNote.pdf

6. Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, & Stephan DA (2007). SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. **Bioinformatics** 23 (1), 57–63.

7. Sentrix® HumanHap550 Genotyping BeadChip http://www.illumina.com/downloads/HUMANHAP550_DataSheet.pdf

8. International HapMap Consortium (2005). A haplotype map of the human genome. **Nature** 437: 1299-1319.

9. Conrad DF, Andrews TD, Carter NP, Hurles ME, & Pritchard JK (2006). A high-resolution survey of deletion polymorphisms in the human genome. **Nature Genetics** 38: 75-81.

10. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, & Altshuler DM (International HapMap Consortium) (2006). Common deletion polymorphisms in the human genome. **Nature Genetics** 39: 86-91.

11. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, & Lee C (2006). Copy number variation: New insights in genome diversity. **Genome Research** 16: 949-961

12. Infinium Genotyping Data Analysis http://www.illumina.com/downloads/GTDataAnalysis_TechNote.pdf

13. Press S.J. Subjective and Objective Bayesian Statistics: Principles, Models, and Applications. 2nd edition. A John Wiley & Sons, Inc.: New Jersey, USA, 2003.

14. The International Hapmap Consortium (2003). The International HapMap Project. **Nature** 426: 789-796

15. Day N. , Genes, Environment and Substance Use and Abuse, text of grant proposal submitted to NIH, 15 March 2007, with permission.

16. Piantadosi S. Clinical Trials: A Methodologic Perspective. 2nd edition. A John Wiley & Sons, Inc.: New Jersey, USA, 2005.