4774–4782 *Nucleic Acids Research, 2012, Vol. 40, No. 11* *Published online 22 February 2012*
doi:10.1093/nar/gks169

Unrecognized sequence homologies may confound genome-wide association studies

Pierre Galichon[1,2,3,*], Laurent Mesnard[1,2,3], Alexandre Hertig[1,2,3], Bénédicte Stengel[4] and Eric Rondeau[1,2,3]

[1]INSERM UMR S702, [2]Université Pierre et Marie Curie – Paris 6, 75006 Paris, [3]Urgences Néphrologiques et Transplantation Rénale, Hôpital Tenon, Assistance Publique des Hôpitaux de Paris, 75020 Paris and [4]INSERM UMR S1018, 94807 Villejuif, France

Received October 12, 2011; Revised January 17, 2012; Accepted January 21, 2012

ABSTRACT

Genome-wide association studies (GWAS) have become a preferred method to identify new genetic susceptibility loci. This technique aims to understanding the molecular etiology of common diseases, but in many cases, it has led to the identification of loci with no obvious biological relevance. Herein, we show that previously unrecognized sequence homologies have caused single-nucleotide polymorphism (SNP) microarrays to incorrectly associate a phenotype to a given locus when in fact the linkage is to another distant locus. Using genetic differences between male and female subjects as a model to study the effect of one specific genomic region on the whole SNP microarray, we provide strong evidence that the use of standard methods for GWAS can be misleading. We suggest a new systematic quality control step in the biological interpretation of previous and future GWAS.

INTRODUCTION

Genome-wide association studies (GWAS) use microarrays of oligonucleotide probes to identify associations between single-nucleotide polymorphisms (SNPs) and a given phenotype. DNA is digested by restriction enzymes into restriction fragments of hundreds of bases, marked with fluorescent bases and hybridized on microarrays containing millions of oligonucleotidic probes that are complementary to the SNP's flanking sequences. When a given variant of a SNP is present, the restriction fragment containing it will hybridize on the corresponding probe through the complementarity of the probe and the SNP's flanking sequence, and the variant will be detectable by its fluorecence signal. In theory, a sequence variant with an effect on the phenotype should be located in the region surrounding the identified SNPs. Currently, the interpretation of GWAS is focused on the exploration of these regions (1). In some cases, this strategy has allowed for the discovery of the underlying molecular mechanism of a phenotype or disease (2). However, many SNPs identified to date have not provided physiological insights (1,3). Because these SNPs have been identified with a high level of statistical significance and often have been validated by independent replication studies (1), we believe that they correspond to true differences in the DNA samples used for analysis, and we sought for different reasons for a statistical link between a SNP and a phenotype.

One possible explanation is that we cannot yet comprehend the biological function of the variants we detect. In a recent study, the genetic variations causing the association of a locus with a chronic renal disease was discovered only years after the locus was identified by GWAS (4).

Another and more troublesome possibility is that the SNP microarray technique used for GWAS systematically associates a phenotype with an irrelevant locus, distant from the genetic sequence(s) responsible for the phenotype. This would mean that variations in DNA, although spatially unrelated to the SNP, can alter its corresponding signal on a microarray.

Genetic differences between sexes (i.e. the presence of a Y or a second X chromosome) present the possibility of an experimental design to investigate the effect of a defined chromosome on the whole SNP microarray results, including results concerning autosomes that 'should not' be altered by differences of sex. Therefore, we performed a GWAS on control patients from available data sets, searching for autosomal SNPs associated with sex status that would not be found if the probes on the array are really specific.

*To whom correspondence should be addressed. Tel: +33 1 56 01 83 17; Fax: +33 1 56 01 66 59; Email: galichon@orange.fr

MATERIALS AND METHODS

Data sets

Five different data sets from previous publications were used. Data set 1 was obtained from 161 control subjects (45 male and 116 female subjects) using the Illumina Quad v3 370 k microarray (5). Data set 2 was obtained from 126 control subjects (64 male and 62 female subjects) using the Affymetrix 500 k array (6). Data set 3 was obtained from the HapMap CEU phase 2 and included 90 subjects (44 male and 46 female subjects). Data set 4 was obtained from the HapMap CEU phase 3 and included 165 subjects (80 male and 85 female subjects) (7). Data set 5 was obtained from 100 control subjects (50 male and 50 female subjects) using an Affymetrix 6.0 microarray (8).

Statistical analysis

Associations and correlations were considered statistically significant when the *P*-value was $<10^{-7}$. No filters were set for Hardy–Weinberg equilibrium, minor allele frequency, or no-call rate, as our study uses sex difference to study the effect of sex chromosome variations, which do not follow the same distribution as autosomes. For the analysis of Data sets 1 through 4, we performed association test on sex using the PLINK whole genome association analysis toolset (Purcell, PLINK v1.07, http://pngu.mgh.harvard.edu/purcell/plink/) (9). For the analysis of Data set 5, we compared the probe intensities in men versus women using a two-sided *t*-test with the MutiTtest function of the ClassComparison package for the R software (Coombes, htpp://bioinformatics.mdanderson.org/OOMPA, Team, R Development Core, http://www.R-project.org/). Next, we calculated the average intensity value for the most reproducible SNPs (intensity values of replicate probes of a SNP showing a correlation with $r > 0.7$ and $P < 10^{-9}$ by Pearson's linear correlation test) and analyzed the correlation between these 46 SNPs' average probe intensity ratios in female subjects using a Pearson's linear correlation test with the two-sided correlation test function of *R*.

Sequence alignments

The Basic Local Alignment Search Tool (BLAST) (10) was used to search for sequence alignments on the human genome in the NCBI build 37.2. For short sequences (probes), the search parameters were set to default for the BlastN algorithm except a word size of 15, an expect threshold of 0.05, and no filter for low complexity and species-specific repeats. For larger sequences (restriction fragments), the search parameters were set to default for the megablast algorithm except a word size of 20 and an expect threshold of 0.05.

Identification of studies with false results

We performed a literature-wide search for GWAS that identified the genes neighboring the SNPs. We searched PubMed and the Gwascatalog (www.genome.gov/gwastudies, accessed 1 December 2011) for genes neighboring the gender-associated SNPs. We picked a few studies for which the precise data needed to check the hypothesis of a sex-related bias were available to us, and we analyzed the data in detail.

RESULTS

Sex modifies the results for autosomal SNPs in microarrays

We performed a genome-wide association study on sex in four independent data sets. All data were from control subjects. The data were obtained using the following technologies: an Affymetrix 500 k microarray, an Illumina 370 k microarray, and HapMap CEU phase 2 and phase 3 genotypes. In all four data sets, we found SNPs that were allegedly located on autosomes but that exhibited significantly different genotype frequencies in men and women. These results are detailed in Table 1. When the analysis was restricted to highly statistically significant SNPs ($P < 10^{-7}$), we were still able to identify six SNPs from the Affymetrix 500 k array and six SNPs from the Illumina 370 k array that were located on autosomes and associated with sex. The analysis of the HapMap data yielded 35 and 17 SNPs from the HapMap phase 2 and HapMap phase 3 genotypes, respectively. Interestingly, one locus was associated with sex in all the data sets (near the TPTE2 gene), and four other loci were found in at least two datatsets (near the WWC2/CDKN2AIP, ADAMTSL3/UBE2QP1, PPP1R12B and PTGER4 genes).

Replicated sequences in autosomes and sex chromosomes explain the effect of sex on autosomal SNPs

Because Mendelian principles of allelic transmission do not explain the association of autosomal loci with sex, we investigated whether nucleotide sequences on sex chromosomes could hybridize to the oligonucleotide probes of autosomal SNPs in various microarrays. Analysis of 28 of the SNP-flanking sequences (i.e. one for each autosomal locus we had found associated with sex in the first step) using the BLAST revealed that 21 of the 28 probes shared total or partial homology with sequences on the Y or X chromosome. All alignments of the SNP-flanking sequences and their locations on the genome can be found in Supplementary Data sets S1 and S2. Figure 1 shows the sequence alignment of a representative SNP-flanking sequence with an autosomal target sequence and with the homolog on a sex chromosome. We picked 28 random SNPs among those who were not found to be associated with sex and used them as control. The BLAST alignment showed that 26 of 28 SNPs had flanking sequences fully specific of their theoretical location, one had one homology on another autosome, and only 1 of 28 had many weak homologies on other chromosomes including chromosome X (Supplementary Data set S3). We then aligned all probes' flanking sequences from Data sets 1 and 2 on the chromosome X and Y sequence, and the association of autosomal SNPs with sex versus homologies on sex chromosomes is represented in Figure 2 and Supplementary Figure S1. When comparing Chi square statistics of probes with homologies versus probes with no homologies on sex chromosomes,

Table 1. SNPs with genotypes significantly associated with gender according to various platforms

rsID	*Chromosome*	*Position*	Genes	SNP	Odds ratio (female/male)	*P*-value
Affymetrix 500 k						
rs4862188	4	184730519	WWC2/CDKN2AIP	T/C	0	7.775×10^{-20}
rs2880301	13	18998534	**TPTE2/MPHOSPH8**	T/C	0	8.685×10^{-20}
rs3883013	15	82889661	**ADAMTSL3/ZSCAN2/UBE2QP1**	C/T	0	8.685×10^{-20}
rs3883011	15	82889398	**ADAMTSL3/ZSCAN2/UBE2QP1**	G/C	0	1.368×10^{-19}
rs3883014	15	82889733	**ADAMTSL3/ZSCAN2/UBE2QP1**	C/G	0	1.527×10^{-19}
rs2228276	19	63271452	ZNF 773/ZNF135	T/C	37.3	9.515×10^{-08}
Illumina 370 k						
rs12734338	1	200736346	**PPP1R12B**	C/T	0	3.57×10^{-31}
rs3881953	1	200794644	**PPP1R12B**	A/G	0	4.26×10^{-31}
rs3817222	1	200731383	**PPP1R12B**	T/C	0	1.02×10^{-30}
rs12743401	1	200743271	**PPP1R12B**	C/T	0	1.02×10^{-30}
rs34868670	5	40273600	PTGER4	C/T	0	1.53×10^{-30}
rs2451078	13	18996289	**TPTE2**	G/C	0.03111	8.56×10^{-25}
Hapmap CEU v2						
rs1556557	1	241046639	RSL24D1P4, LOC10012	A/G	0	6.06×10^{-15}
rs3817227	1	200731465	**PPP1R12B**	G/A	0	6.06×10^{-15}
rs4084639	1	200776787	**PPP1R12B**	C/G	0	6.06×10^{-15}
rs10914658	1	33303337	AK2A, ADC	A/G	0	6.06×10^{-15}
rs12734001	1	200657537	**PPP1R12B**	T/C	0	6.06×10^{-15}
rs12739153	1	241049487	RSL24D1P4, LOC10012	T/G	0	6.06×10^{-15}
rs12741415	1	200741397	**PPP1R12B**	A/G	0	6.06×10^{-15}
rs17319010	1	222156006	ACTBP11, CIPC5	C/A	0	6.06×10^{-15}
rs17802433	2	94901357	TEKT4	T/G	0	6.06×10^{-15}
rs4862188	4	184592364	LOC100127981, CDKN2AIP	T/C	0	6.06×10^{-15}
rs2999200	13	18887941	**TPTE2**	T/C	0	6.06×10^{-15}
rs3883011	15	82889398	**UBE2Q2P1**	C/G	0	6.06×10^{-15}
rs3883013	15	82889661	**UBE2Q2P1**	C/T	0	6.06×10^{-15}
rs17301021	15	82613080	**ADAMTSL3**	G/C	0	6.06×10^{-15}
rs2502344	1	241137354	LOC100129949, LOC100420263	A/G	0	6.81×10^{-15}
rs12734338	1	200736346	**PPP1R12B**	C/T	0	6.81×10^{-15}
rs3883014	15	82889733	**UBE2Q2P1**	G/C	0	6.81×10^{-15}
rs3881953	1	200794644	**PPP1R12B**	A/G	0	7.67×10^{-15}
rs1778596	1	143702635	**PDE4DIP**	A/T	0	8.66×10^{-15}
rs12743401	1	200743271	**PPP1R12B**	C/T	0	8.66×10^{-15}
rs2880301	13	18998534	**TPTE2**	T/C	0	1.06×10^{-14}
rs3847124	7	137842064	TRIM24	G/A	0	1.53×10^{-14}
rs11166266	1	99771825	LPPR4, PALMD	T/C	0	1.87×10^{-14}
rs12723357	1	241185135	LOC100129949, LOC100420263	C/T	0	1.87×10^{-14}
rs3013398	1	241209589	LOC100129949, LOC100420263	T/C	87	8.60×10^{-14}
rs2390647	1	91130771	LOC100505821, ZNF644	C/T	∞	9.65×10^{-14}
rs17042395	3	16568435	**RFTN1**	G/A	0.01149	1.09×10^{-13}
rs12372818	13	46581126	HT2RA	A/G	0.02222	1.93×10^{-13}
rs351881	20	62314104	MYT1	T/C	0.02222	2.19×10^{-13}
rs6820128	4	91700109	FAM190A	A/G	0	1.13×10^{-12}
rs11667496	19	23750678	RPSAP58	G/A	0.01266	1.45×10^{-12}
rs4860568	4	64690977	TECRL	A/G	0.01299	2.03×10^{-12}
rs9881157	3	35626953	ARPP21	C/A	0.02564	1.92×10^{-11}
rs4685345	3	16585452	**RFTN1**	G/C	0.099	6.16×10^{-10}
rs6803924	3	16592069	**RFTN1**	G/C	0.09702	1.51×10^{-09}
Hapmap CEU v3						
rs34868670	5	40273600	PTGER4	C/T	0	3.631×10^{-26}
rs4737118	8	43533172	**POTEA**	G/A	0	3.631×10^{-26}
rs12743401	1	200743271	**PPP1R12B**	C/T	0	4.603×10^{-26}
rs12214551	6	2991748	SERPINB8P1	C/T	0	5.635×10^{-26}
rs36019094	5	40273131	PTGER4	A/C	0	8.188×10^{-26}
rs7808552	7	63066168	VN1R36P, LOC100419780	G/A	0	9.278×10^{-26}
rs3817222	1	200731383	**PPP1R12B**	T/C	0	9.839×10^{-26}
rs3994533	15	82882831	**ADAMTSL3, UBE2Q2P1**	T/C	0	9.839×10^{-26}
rs2880301	13	18998534	**TPTE2**	T/C	0	1.359×10^{-25}
rs12741415	1	200741397	**PPP1R12B**	A/G	0	1.763×10^{-25}
rs6944297	7	63937080	ZNF138, LOC168474	T/G	0	2.458×10^{-25}
rs6836144	4	119595470	LOC100128177, LOC100420037	A/C	∞	5.355×10^{-25}
rs1556557	1	241046639	RSL24D1P4, LOC100129949	A/G	0.006211	1.77×10^{-24}
rs7039117	9	97097001	FANCC	C/T	0.006617	2.553×10^{-23}
rs6917603	6	30125050	ETF1P1, C6Orf12	C/T	0.0559	6.801×10^{-20}
rs9636470	2	87947576	LOC730268, LOC100419917	G/A	3.569	3.869×10^{-08}
rs11635160	15	82607789	**ADAMTSL3, UBE2Q2P1**	A/G	0.2805	7.955×10^{-08}

In bold are the SNPs that also were identified in an Affymetrix 6.0 data set by directly comparing probe intensities.

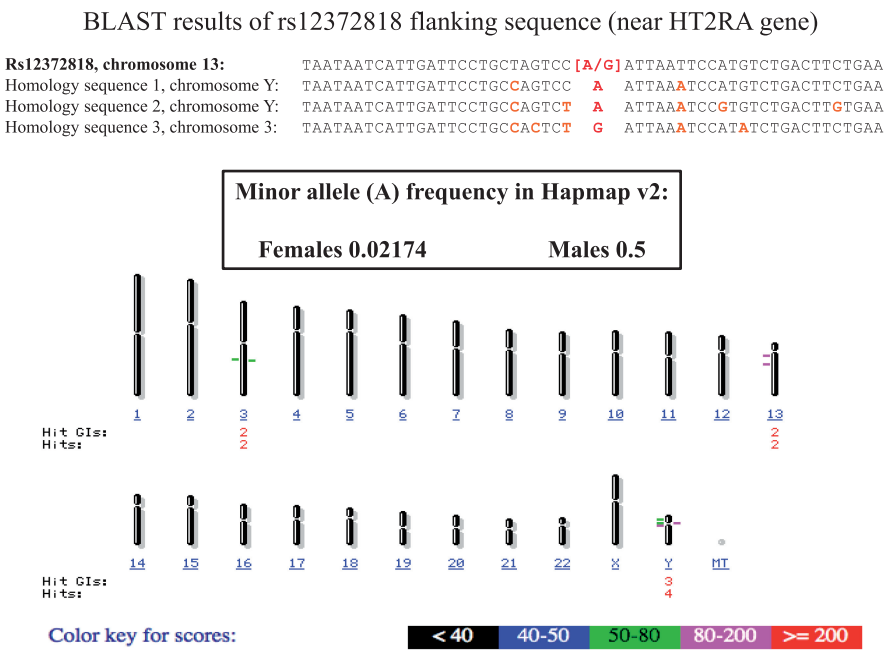

Figure 1. BLAST alignment analysis of the flanking sequence of a sex-associated SNP (rs12372818 on chromosome 13). Two homologous sequences are present on the Y chromosome (and one on chromosome 3). The presence of the 'A' variant on chromosome Y is responsible for a higher frequency of the minor allele in males.

we found that some groups of probes with high homologies on sex chromosomes had a significantly higher association to sex. Interestingly, in Data set 2 (Supplementary Figure S1), this was still true after exclusion of all SNPs showing an association to sex after Bonferroni correction.

The 7 other SNPs did not exhibit such strong homology between their flanking sequences and sex chromosome sequences. We therefore investigated the possibility that a competition would occurr between restriction fragments from the sex chromosomes and autosomal restriction fragments, with respect to hybridization of the oligonucleotide probe.

We used BLAST to search the entire genome for sequences exhibiting homology within the larger region corresponding to the restriction fragment containing the SNP. We found that the restriction fragments from four of the seven SNPs associated with sex had homologous sequences located on the sex chromosomes. All alignments of the SNP-flanking regions and their locations can be found in Supplementary Data sets S4 and S5. Supplementary Data set S6 shows the alignment of a representative SNP restriction fragment with sex DNA.

Schematics of the two mechanisms that we have identified as possibly biasing SNP association results are presented in Figure 3.

The study of probe intensities increases sensitivity in the search for SNP interference

Because no strong sex chromosome homology was found for some SNPs, and because we found that homologies often were present on other autosomes, we suspected that weaker and/or repeated homologies might be sufficient to influence microarray results for some SNPs. To verify this hypothesis and because we were surprised that some loci were associated with sex in some data set and not in others, we decided to use a fifth data set for a more refined analysis.

This time, we analyzed the probe intensity values (rather than the genotype) in relation to the sex status on a fifth data set (Affymetrix 6.0). We found that 126 autosomal SNPs were significantly influenced by sex status (Supplementary Table S1). Remarkably, this intensity-based approach (studying a continuous variable) proved to be very powerful for detecting the influence of sex on these SNPs, as it allowed the identification of twice as many loci from only one data set as had been identified in four different data sets using the genotype-based approach. In addition, these results corroborated the results obtained by comparing genotypes (33 of 64 SNPs identified by the genotype-based approach were located in regions identified by the intensity-based approach). This confirms that most of the associations we observed were neither fortuitous nor specific to a single microarray technology but were relevant to all SNP microarrays.

However, one SNP near PTGER4 was found to be strongly associated with sex in the Illumina data set ($P = 1.53 \times 10^{-30}$), but it was far from significant in the 225 SNPs neighboring PTGER4 in the Affymetrix 6.0 data set (Supplementary Table S2), indicating that the association with sex status was restricted to a single SNP, not the entire locus.

A literature-wide search for these loci allowed to detect and correct errors because of a sex-related bias

We searched for published genotyping studies that had reported the identification of loci containing sex-dependent SNPs. We found that the *post hoc* verification

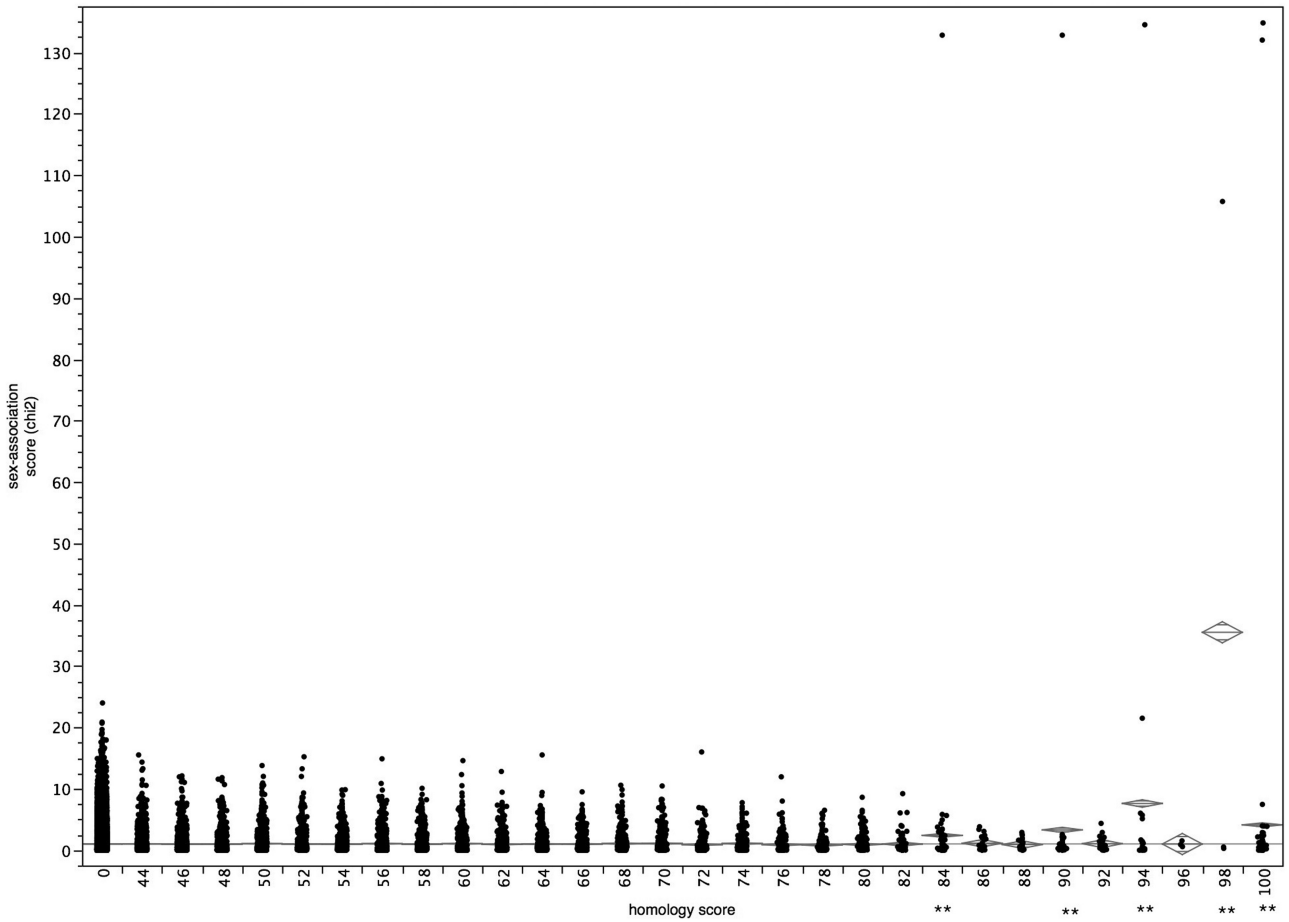

Figure 2. Sex-association score (Chi square statistics) versus the homology score (BLAST raw score) in data set 1 (Illumina 370 k). For each level of homology, mean diamonds with 95% confidence interval. $^{**}P < 0.0001$ versus no homology (0).

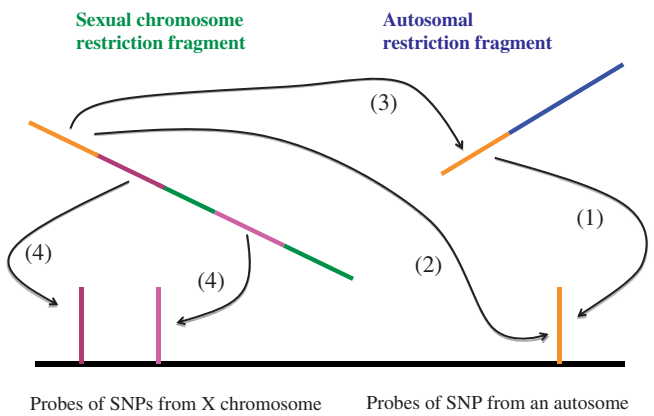

Figure 3. Schematic view of the hybridization of DNA to a microarray probe. Three possibilities include theoretical hybridization, rogue hybridization with a homolog, and bulk hybridization of genomic DNA that sequesters the restriction fragment away from the probe. (1) Hybridization of the target sequence with the probe, according to theory. (2) Hybridization of a sex chromosome sequence with the probe of a homologous autosomal SNP, competing with the theoretical autosomal restriction fragment. (3) Hybridization of a sex chromosome restriction fragment with an autosomal SNP restriction fragment, competing with the microarrays' oligonucleotide probe. (4) Oligonucleotide probes for sex chromosomes' SNPs hybridize with the same restriction fragment as probes for autosomal SNPs and are thus statistically correlated.

of genotyping analysis often was impossible (because of difficulties in getting access to the raw data). Another difficulty stems from the frequent use of imputation to create virtual SNPs from other nearby SNPs, e.g. to merge data from various microarray platforms. This means that a SNP with a flanking sequence duplicated in the genome can be imputed to a virtual SNP with a unique flanking sequence. However, we were able to select three studies in which cryptic duplications of SNP flanking sequences on a sex chromosome have led to the publication of erroneous results:

Study 1. A PPP1R12B allele was found to be preferentially transmitted from parents to offspring, a phenomenon that is called 'transmission distortion' by the authors (11). They note that a SNP near PPP1R12B has an unexpectedly high frequency of heterozygotes when it is transmitted from male parents. In the light of our data showing that PPP1R12B lies in a region duplicated in the Y chromosome causing the SNPs near PPP1R12B to be biased by sex, it is more likely because of the transmission of a 'third PPP1R12B allele' on the Y chromosome from father to son.

Study 2. A SNP near TPTE2 was found to be associated with the presence of hepatocarcinoma in patients with

liver cirrhosis (8). This association was in fact due to an homology of TPTE2 region on the Y chromosome and a sex ratio of 3.4 in hepatocarcinoma versus 1.3 in liver cirrhosis (12). The authors removed TPTE2 from their results after our letter (13).

Study 3. A locus near PTGER4 was found to be associated with multiple sclerosis in a meta-analysis (14). We have found the association of a SNP near PTGER4 with sex in Illumina (because of a sequence homology on the Y chromosome) but not in Affymetrix microarray (c.f. text above, Table 1 and Supplementary Data set S1). This meta-analysis used Illumina data in 37% of the 2624 cases and only 12% of the 7220 controls, indicating that the proportion of males tested on Illumina platform (i.e. subject to the sex-related bias toward PTGER4) was larger in cases than in controls (10 versus 3%). The virtual SNP rs6896969 (obtained by imputation after merging Affymetrix and Illumina data) near PTGER4 was associated (without correction on sex and cohort of origin) with the disease with $P = 10^{-7}$. The authors, because of a strong preponderance of women, performed a second analysis using sex and cohort of origin as covariates to identify additional suceptibility loci for the disease. They publish in Supplementary Data, but do not comment, that the association of rs6896969 with the disease loses genome-wide significance ($P = 10^{-2}$). By joining their analysis (unadjusted on sex and cohort of origin) with the replication analysis, they find a significant association of PTGER4 with the disease, a result they highlight in their conclusion. We think this association is due to the bias we describe here and has no physiopathological significance.

Replicated sequences modify the results of SNPs regardless of sex

We next asked whether replication of an autosomal sequence containing a SNP on another autosome could influence the microarray result concerning that SNP. This is especially important as the filters usually used on the data set (e.g. Hardy–Weinberg equilibrium, no-call rate) would be less likely to eliminate SNPs with interautosomal homologies than SNPs with homologies on sex chromosomes. As these replicated sequences could be present anywhere in the genome, it would take $\sim 5 \times 10^{11}$ correlation calculations to investigate all possible combinations. This analysis would require both more computational power than we have and more patients to achieve the statistical power required to take the necessary multiple test correction into account (15). Instead, we chose to study the correlation of autosomal SNPs that were associated with sex in the first step of our study. We chose to study these SNPs in women, as they have two X chromosomes and, thus, a SNP distribution similar to autosomes. This provided us with a model of interautosomal SNP correlation in which we had only a handful of SNPs to test, preselected for their high probability to beeing influenced by chromosome X. We looked for significant correlations between a selection of 46 autosomal SNPs we had found to be influenced by gender status and any of the SNPs located on the X chromosome. We found that 31 autosomal SNPs (67%) were significantly correlated with at least one (but up to 10^3) SNPs on the X chromosome (Supplementary Data set S7). We used BLAST to search for alignment of the X chromosome with either the sequences of the autosomal SNPs' probes or the SNPs' restriction fragments. Thus, we identified repetitive homologies in loci from the X chromosome in locations where we had found SNPs with significant correlations with autosomal SNPs. Interestingly, we found that some SNPs could be significantly correlated even when only short homologous sequences were involved (Table 2).

DISCUSSION

Since the first GWAS using SNP microarrays, the reality of discoveries using this method has been the subject of intense debate (2,16). Although such an unbiased, systematic genome-wide approach is very appealing, technically, this approach consists in looking for a needle in a haystack without knowing what the needle looks like. Here, we found that SNP microarrays, although they varied in design and were performed on different individuals, yield reproducible information that correspond to true biological properties. Our independent association studies on gender repeatedly highlighted SNPs related to the sex chromosomes by sequence homology. However, our findings also demonstrate that, to date, technical flaws pertaining to SNP microarrays have occurred, affecting the information that they are designed to retrieve. Although it can be expected that the association of a SNP with a given phenotype will reflect a molecular mechanism involving the single genomic region surrounding that very SNP, it actually integrates many interactions between more-or-less homologous sequences also subject to variations but without any relevance with respect to the studied locus. Our results show that, in four separate data sets obtained from various genotyping platforms, some SNPs systematically give spurious results. Although these homologous sequences are easily detectable when they are located on sex chromosomes, they are not systematically eliminated, which exposes to the posiibility of misleading findings. We have verified that our results have practical applications in GWAS, showing that this bias has led the authors of these studies to identify statistical associations of SNPs with a phenotype with no underlying biological relevance (8,11,14).

We also demonstrate that homologies between two autosomal regions cause errors that may be both more frequent and more cryptic. Thus, extending our findings on sex chromosomes to the whole genome should detect other yet unrecognized homologies. Overall, the conditions of our analysis, which was performed on a limited number of subjects and investigated effects because of sex chromosomes only, suggest that the actual number of SNPs that confer a bias and jeopardize the interpretation of the results might be much greater.

The presence of artifactual results in microarrays has been predicted in previous publications. In Musumeci's

Table 2. Example on SNP rs13269433 of convergent approaches using BLAST sequence alignment to identify interautosomal SNP homologies and a correlation test to identify interdependent SNPs

rs13269433, chromosome 8, near MFHAS1					
Flanking sequence: ATATATATCAGCCAGA[T/C]GTGCCACGTGAGCCTG					
Blast hits	Alignment	Position on chromosome X	rsID	Correlation (r)	Correlation (P)
Haloacid dehalogenoase-like hydrolase domain-containing protein	ATATATATCAGCCA	245274	rs12007101	−0.79	1.37×10^{-11}
			rs5934477	0.74	5.94×10^{-10}
Mastermind-like domain-containing protein 1	TGCCACGTGAGCCT	551801	rs6649480	−0.74	7.47×10^{-10}
			rs9723770	−0.78	1.86×10^{-11}
			rs5925461	−0.74	9.30×10^{-10}
			rs5970516	−0.85	9.54×10^{-15}
			rs5925482	0.82	4.56×10^{-13}
Kelch-like protein 13	TATATCAGCCAGA	777982	rs10465428	0.75	3.37×10^{-10}
			rs7885432	−0.80	5.59×10^{-12}
			rs2465941	0.80	4.66×10^{-12}
			rs2106683	0.76	1.34×10^{-10}
Neuroligin-4. X-linked precursor	ATATATATCAGCCA	922956	rs17219044	0.75	5.18×10^{-10}
			rs36122347	0.76	2.13×10^{-10}
			rs16983683	−0.80	4.95×10^{-12}
			rs5961738	0.84	1.75×10^{-14}
			rs12844412	−0.76	1.15×10^{-10}
			rs7881412	0.75	4.01×10^{-10}
			rs10127411	0.80	4.35×10^{-12}
DDB1- and CUL4-sassociated factor 12-like protein 1	ATATATCAGCCAGA	1369645	rs5929972	−0.80	1.95×10^{-12}
			rs7065014	−0.74	9.58×10^{-10}
			rs201647	−0.77	9.61×10^{-11}
			rs1601226	0.83	1.57×10^{-13}
			rs16997689	0.80	4.09×10^{-12}
PAS domain-containing protein 1	ATATATATCAGC	1588059	rs16995984	0.78	2.18×10^{-11}
			rs7051678	0.74	5.78×10^{-10}
			rs5924663	0.78	2.56×10^{-11}
Gamma-aminobutyric acid receptor subunit alpha 3 precursor	TATATATCAGCCA	2591595	rs7057635	−0.75	4.41×10^{-10}
			rs4446880	0.75	2.88×10^{-10}
Ribose-phosphate pyrophosphokinase 2	TATATATCAGCCA	4384225	rs16987131	0.75	3.69×10^{-10}
Nance–Horan syndrome protein isoform 1	CCACGTGAGCCTG	9035432	rs7887450	0.80	4.72×10^{-12}
			rs6632979	−0.76	1.09×10^{-10}
			rs7473191	0.83	1.06×10^{-13}
			rs6527811	−0.81	1.14×10^{-12}
Dystrophin	TATATATCAGCCA	24534293	rs16989676	0.77	8.99×10^{-11}
			rs16989902	−0.77	6.69×10^{-11}
			rs1158629	0.75	2.40×10^{-10}
			rs1356619	0.75	3.07×10^{-10}
			rs1518519	−0.82	3.63×10^{-13}
			rs7887670	−0.74	8.15×10^{-10}
Melanoma-associated antigen B16	TGCCACGTGAGCCT	27045018	rs6632359	0.80	1.26×10^{-12}
Zinc finger protein 92 homolog	TGCCACGTGAGC	152706248	rs2980024	−0.79	1.28×10^{-11}

From left to right, for each line, the gene nearest to BLAST hit (region of homology to rs13269433 on chromosome X), the aligned sequence, its position on chromosome X, the correlated SNPs in the same region, its correlation factor r and its P-value.

in silico study, the presence of duplicated sequences with a single nucleotide difference was estimated to represent 8.3% of all SNPs from the dbSNP database (17). The authors argued that these duplicated sequences polluted microarrays with SNPs that could never be associated with the studied phenotype. In a more recent study, Doron and Shweiki (18) show that 11.9% of Hapmap SNPs align to the genome non-uniquely (30 nt's upstream and downstream to SNP position). They suggest that the SNP uniqueness problem is a potentially massive bias in genotyping analysis. Here, we show that it indeed leads to false-positive associations. We show that replicated sequences actually can be responsible for the identification of false associations. Furthermore, we show that even weak homologies can modify the microarray results. These data corroborate the experimental results of Eklund, who showed that even weak similarities are sufficient to bias microarray probes when 10% of hemoglobin cDNA is added to the chip (19). In our study, we use a genome-wide approach, focusing our study on microarray data. However, the bias we discovered is not specific to microarray technology but could occur in other types of genotyping study.

The genome is known to be rich in repetitive sequences (20). Most of these sequences are considered to be 'junk DNA' because they have no functional promoter regions

and are not expressed. Sex chromosomes contain large amounts of these repetitive sequences (21–23). The telomeric regions are especially rich in repetitive sequences and are especially prone to neomutation (24). Thus, a special attention should be paid to these repetitive sequences when studying a pathological trait (24), and duplications should be taken into consideration when interpreting GWAS.

Studying genotypes can point at true statistically relevant association between marker and traits, which cannot be sorted out by increasing sample sizes (25). Our study shows that cryptic sequence duplication can cause such indirect association between markers and traits, but we believe that our results may help microarray constructors and bioinformatics specialists to improve the design of array chips and the processing of their results to make GWAS more reliable. SNPs with flanking sequences that are not specific to a single genomic region should be replaced every time a SNP with specific flanking sequences exists within the same region (17,18). At minimum, microarray constructors should clearly mention this ambiguity in their annotation file. Our guess is that the SNPs have not been updated since the completion of human genome sequencing (26), and the selection of these misleading SNPs might have promoted their high level of apparent heterozygosis. As sequence duplication are frequent in the genome, the risk of including a SNP with duplicated flanking sequence is high (17,18). Lastly, it might be that some of these duplications did not exist in the populations where the SNPs were first reported. The discrepancy we found between the Illumina and Affymetrix microarrays concerning the PTGER4 bias not only indicates that the bias 'can' be avoided, at least in some cases, but also stresses the difficulty of pooling data obtained from various microarray platforms.

In the meantime, we recommend that the interpretation of previous and future GWAS be reconsidered in light of our findings. Errors in GWAS results caused by repetitive sequences can be avoided by several means. The usual exclusion tests (no-call SNPs, SNPs with low minor allele frequency, and SNPs not matching Hardy–Weinberg equilibrium) are useful, but applying them more strictly might eliminate SNPs that are strongly influenced by natural selection (27). Instead, we suggest three steps that provide more confidence in the results without excluding SNPs from the analysis. First, stratification based on sex and on the platform used for genotyping should be performed systematically, even if it could diminish the statistical power of the analysis (27,28). Second, the specificity of all identified SNP sequences should be systematically checked. This should include a genome-wide alignment of the SNP-flanking sequences and of the restriction fragments. If significant homologies are found in other genomic regions, these should be considered as susceptibility loci as well. However, we found that, in some cases, the effect of replicated sequences on SNP results is difficult to predict by sequence alignment, especially when the homologies are weak. Third, the full sequencing of the susceptibility loci associated with one SNP should help identify which of the replicated sequences is truly associated with the phenotype.

In sum, our findings underscore the need for a very thoughtful analysis of SNPs associated with a phenotype to discriminate misleading data devoid of any biological relevance. We would like to stress that the raw data from previously published studies should be available to the scientific community. In practice, external access to data for verification purposes is difficult, delayed and sometimes denied (although data are duly referenced in the dbGAP database) (29,30). We urge authors who have reported strong statistical associations of SNPs with diseases to perform a secondary analysis. Some SNPs that are not surrounded by any relevant gene with respect to a specific disease may have been selected because of their duplication on sex chromosomes or even on autosomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figure 1 and Supplementary Data sets 1–7.

ACKNOWLEDGEMENTS

We thank Dr Bertrand Jordan for critical discussion and reading of the manuscript.

FUNDING

P.G. recieived a scholarship from the Société Francophone de Transplantation. Funding for open access charge: INSERM.

Conflict of interest statement. None declared.

REFERENCES

1. Manolio,T.A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, **363**, 166–176.
2. Hirschhorn,J.N. (2009) Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699–1701.
3. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
4. Genovese,G., Friedman,D.J., Ross,M.D., Lecordier,L., Uzureau,P., Freedman,B.I., Bowden,D.W., Langefeld,C.D., Oleksyk,T.K., Uscinski Knob,A.L. *et al.* (2010) Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, **329**, 841–845.
5. Stanescu,H.C., Arcos-Burgos,M., Medlar,A., Bockenhauer,D., Kottgen,A., Dragomirescu,L., Voinescu,C., Patel,N., Pearce,K., Hubank,M. *et al.* (2011) Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N. Engl. J. Med.*, **364**, 616–626.
6. Hafler,D.A., Compston,A., Sawcer,S., Lander,E.S., Daly,M.J., De Jager,P.L., de Bakker,P.I., Gabriel,S.B., Mirel,D.B., Ivinson,A.J. *et al.* (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.*, **357**, 851–862.
7. Thorisson,G.A., Smith,A.V., Krishnan,L. and Stein,L.D. (2005) The International HapMap Project Web site. *Genome Res.*, **15**, 1592–1593.

8. Clifford,R.J., Zhang,J., Meerzaman,D.M., Lyu,M.S., Hu,Y., Cultraro,C.M., Finney,R.P., Kelley,J.M., Efroni,S., Greenblum,S.I. *et al.* (2010) Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma. *Hepatology*, **52**, 2034–2043.
9. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
10. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Deng,L., Zhang,D., Richards,E., Tang,X., Fang,J., Long,F. and Wang,Y. (2009) Constructing an initial map of transmission distortion based on high density HapMap SNPs across the human autosomes. *J. Genet. Genomics*, **36**, 703–709.
12. Galichon,P., Hertig,A., Rondeau,E. and Mesnard,L. (2011) Warning: genome-wide association studies can be misleading. An example in hepatology. *Hepatology*, **53**, 1408.
13. Clifford,R.J. and Buetow,K.H. (2011) Reply to galichon, *et al* *Hepatology*, **53**, 1408–1409.
14. De Jager,P.L., Jia,X., Wang,J., de Bakker,P.I., Ottoboni,L., Aggarwal,N.T., Piccio,L., Raychaudhuri,S., Tran,D., Aubin,C. *et al.* (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.*, **41**, 776–782.
15. Cantor,R.M., Lange,K. and Sinsheimer,J.S. (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **86**, 6–22.
16. Altmüller,J., Palmer,L.J., Fischer,G., Scherb,H. and Wjst,M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.*, **69**, 936–950.
17. Musumeci,L., Arthur,J.W., Cheung,F.S., Hoque,A., Lippman,S. and Reichardt,J.K. (2011) Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.*, **31**, 67–73.
18. Doron,S. and Shweiki,D. (2011) SNP uniqueness problem: a proof-of-principle in HapMap SNPs. *Hum. Mutat.*, **32**, 355–357.
19. Eklund,A.C., Friis,P., Wernersson,R. and Szallasi,Z. (2010) Optimization of the BLASTN substitution matrix for prediction of non-specific DNA microarray hybridization. *Nucleic Acids Res.*, **38**, e27.
20. Jurka,J., Kapitonov,V.V., Kohany,O. and Jurka,M.V. (2007) Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.*, **8**, 241–259.
21. Kondo,M., Hornung,U., Nanda,I., Imai,S., Sasaki,T., Shimizu,A., Asakawa,S., Hori,H., Schmid,M., Shimizu,N. *et al.* (2006) Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. *Genome Res.*, **16**, 815–826.
22. Makova,K.D., Yang,S. and Chiaromonte,F. (2004) Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res.*, **14**, 567–573.
23. Matsunaga,S. (2009) Junk DNA promotes sex chromosome evolution. *Heredity*, **102**, 525–526.
24. Mazzarella,R. and Schlessinger,D. (1998) Pathological consequences of sequence duplications in the human genome. *Genome Res.*, **8**, 1007–1021.
25. Platt,A., Vilhjálmsson,B.J. and Nordborg,M. (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics*, **186**, 1045–1052.
26. Durbin,R.M., Abecasis,G.R., Altshuler,D.L., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
27. Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781–791.
28. Ziegler,A., König,I.R. and Thompson,J.R. (2008) Biostatistical aspects of genome-wide association studies. *Biom. J.*, **50**, 8–28.
29. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
30. Ochsner,S.A., Steffen,D.L., Stoeckert,C.J. and McKenna,N.J. (2008) Much room for improvement in deposition rates of expression microarray data sets. *Nat. Methods*, **5**, 991.