Genome-wide haplotyping of single cells

Thierry Voet,[1] Evelyne Vanneste,[1,2] Hilde Peeters,[1] Peter Konings,[3] Yves Moreau,[3] Jean-Pierre Fryns,[1] Thomas D'Hooghe,[2] Eric Legius,[1] and Joris R. Vermeesch[1,4]

[1]Center for Human Genetics, K.U.Leuven, Herestraat 49, 3000 Leuven, Belgium; [2]Leuven University Fertility Center, University Hospital Gasthuisberg, Herestraat 49, 3000 Leuven, Belgium; [3]ESAT-SISTA, K.U.Leuven, Kasteelpark Arenberg 10 , 3001 Heverlee, Belgium

[4]Corresponding author.

Center for Human Genetics, University Hospital Gasthuisberg, Herestraat 49 box 602, B-3000 Leuven, Belgium. E-mail: Joris.Vermeesch@uz.kuleuven.be; tel. +32 16 345941; fax +32 16 346060

Running title: Genome-wide haplotyping of single cells

Keywords: single cell, genotype, haplotype, homologous recombination, embryo selection

Abstract

Haplotyping has proven invaluable to identify genetic variants associated with simple and complex traits, to study population history as well as to unravel meiotic and mitotic recombination processes. Genome-wide haplotyping methods use DNA samples extracted from a large number of cells. Here, we developed a generic approach to type genome-wide single nucleotide polymorphisms in single human cells and to reconstruct genome-wide haplotypes of single- or dual-cell derived genotypes. Genome-wide sequences of syntenic alleles were determined for EBV-transformed lymphoblastoid cells as well as human blastomeres. To this end, multiple displacement amplified DNA samples of single cells were hybridized to Affymetrix 250K SNP-arrays. Different algorithmic designs were subsequently developed to assess from the single-cell derived SNP-probe intensities the sequence of syntenic alleles and to pinpoint accurately the majority of parental homologous recombination sites across the entire genome using a linkage-based approach. The ability to genome-wide haplotype single cells will open up new research applications to further scrutinize drivers of haplotype diversity, mainly meiotic homologous recombination as well as somatic (homologous) recombination processes that occur often during (human) tumorigenesis. In addition, this method is likely to revolutionize genetic diagnosis of preimplantation *in vitro* fertilized embryos both in human reproduction as well as in animal breeding programs by enabling genome-wide quantitative trait loci selection at the embryonic level.

The SNP-array data has been deposited in NCBI's Gene Expression Omnibus (GEO) and is accessible through GEO Series accession number GSE17318 (http://www.ncbi.nlm.nih.gov/projects/geo/). The supplementary data can be found online (http://genome.cshlp.org/).

Introduction

Haplotyping is a cornerstone of many disciplines in genetics (Buntjer et al. 2005; Garrigan and Hammer 2006; Mackay and Powell 2007; Altshuler et al. 2008; Goddard and Hayes 2009). A haplotype represents the succession of alleles in synteny and can be determined *in silico* from observed genotype data by statistical and computational methods (Gao et al. 2009). The inference of haplotypes has proven invaluable not only to identify genetic variants underlying an inherited disease (Altshuler et al. 2008), but also to unravel the genetic architecture of Mendelian and complex traits that are particularly important for animal and plant breeding programs (Buntjer et al. 2005; Mackay and Powell 2007; Goddard and Hayes 2009), to study evolution and population history (Garrigan and Hammer, 2006) as well as meiotic and mitotic recombination processes (De Raedt et al. 2006; Kotzot 2008; Webb et al. 2008). In addition, computation of the sequence of syntenic alleles can be used to identify genotyping errors by looking for successive homologous recombination sites in a limited locus (Abecasis et al. 2002) and to infer missing genotype calls as well (Burdick et al. 2006). Until recently, genotypes were mainly determined by analyzing one or a few polymorphic markers per reaction. The recent development of several genome-wide high-throughput approaches for marker analyses have allowed to type thousands of polymorphic markers in a single experiment (McCarroll et al. 2008; Beaudet and Belmont 2008). Microarrays with oligonucleotides targeting single nucleotide polymorphisms (SNPs) over the entire genome have enabled massive parallel SNP-typing and have accelerated the identification of genetic variants underlying specific traits or diseases in both humans and animals mainly through linkage analyses and genome-wide association studies (Altshuler et al. 2008; Goddard and Hayes 2009; Frazer et al. 2009). In addition, SNP-typing has been used to identify regions of copy neutral loss of heterozygosity (CN-LOH) on a genome-wide scale in a large series of acute myeloid leukaemia samples, which so far has not been possible through any other technology (Dutt and Beroukhim 2007; Gupta and Young 2009).

Most genotyping and haplotyping methods however make use of large amounts of DNA and, hence, use DNA that is extracted from a large number of cells of the organism. In many applications though, only a few or even a single cell is available for analysis. In the study of meiotic recombination processes often single haploid sperm cells are genotyped (Jiang et al. 2005).

In preimplantation genetic diagnosis, PCR-based 'simple tandem repeat' (STR) marker analysis is often used to (haplo)type a small locus in a single blastomere (Braude et al. 2002; Renwick

and Ogilvie 2007). In addition, the analysis of a mixture of cells obscures the presence of different subpopulations or cells with a different allelic constitution. For example, cancers acquire cumulative changes and tumor DNA-samples are often contaminated with DNA from normal cells as well. As a consequence analyzing such heterogeneous DNA-samples may lead to loss of valuable genetic information.

Genome-wide SNP-typing as well as DNA copy number variation (CNV) detection has recently become simultaneously feasible in single human cells (Iwamoto et al. 2007; Vanneste et al. 2009a; Treff et al. 2009). We developed a novel array-based method to screen for genome-wide copy number variation and loss of heterozygosity in single cells (Vanneste et al. 2009a). Here, we extend this method with a generic approach to reconstruct genome-wide haplotypes of single or few human cells. Proof-of-principle is provided for multiple displacement amplified (MDA)-DNA of single human Epstein-Barr virus (EBV)-transformed lymphoblastoid cells and blastomeres derived from human *in vitro* fertilized (IVF) embryos. In addition, we show that disease-causing alleles can be detected in single human blastomeres via a linkage approach.

Results

Genotyping of single cells

The method was established using four EBV-transformed lymphoblastoid cell lines derived from four different individuals respectively (see Methods). Genomic DNA isolated from multiple blood cells of these individuals as well as MDA-DNA from three single cells per cell line were hybridized to Affymetrix 250K SNP-arrays. The obtained SNP-probe intensities were interpreted by three different algorithms - Dynamic Model (Di et al. 2005), BRLMM and Birdseed (Korn et al. 2008) (see Methods) - that are optimized for typing SNPs in diploid genomes not subjected to MDA, allele drop-out (ADO) or preferential amplification (PA).

SNP-typing accuracy of the MDA single-cell DNA was determined by aligning the single-cell genotypes with the reference genotypes obtained from the corresponding non-amplified genomic DNA. Although the different algorithms had minor effects on the concordances of the single-cell homozygous SNPs with the reference SNPs, a significant effect was noted on the concordances of the heterozygous SNPs (Table 1 and Supplementary Fig. 1). The Dynamic Model algorithm outperforms both the BRLMM and Birdseed algorithms for typing heterozygous SNPs in single human cells. Using

a SNP-typing confidence threshold of 0.12 in the Dynamic Model algorithm on average 89.04 % (± 5.0 % SD) of the homozygous SNPs and 91.24 % (± 6.0 % SD) of the heterozygous SNPs in the single cells had a match with the corresponding genotype determined with genomic DNA isolated from multiple cells.

Furthermore, of the SNPs that are heterozygous in the non-amplified genomic DNA samples on average 29.38 % (±11.30 % SD), 12.41 % (±4.81 % SD) and 32.91 % (±6.33 % SD) are homozygous in the corresponding single cells when genotyped with the Dynamic Model (Di et al. 2005), BRLMM or Birdseed (Korn et al. 2008) algorithm respectively. Part of these inconsistencies may be attributed to whole genome amplification artifacts, like allele drop out or preferential amplification of particular alleles of a single cell's genome (Wells et al. 2008). However, it is clear that different statistical algorithms introduce different calling errors.

Single-cell DNA-fingerprints computed by the Dynamic Model were further explored for haplotype reconstruction.

Single- and dual-cell haplotyping

To investigate the possibility for reconstructing haplotypes from single-cell derived genotypes, we first determined for four individuals belonging to three different families, a reference allelic sequence of the maternally inherited autosomes and sex chromosome. To this end, genomic DNA of the four individuals as well as of their parents and maternal grandparents were typed for 250K SNPs and the haplotypes were computed using the MERLIN algorithm (see Methods). Two recombination sites each in a different individual were validated by STR-marker analysis (data not shown). Subsequently, single-cell derived haplotypes were determined by swapping the genotypes of the four individuals with single-cell derived genotypes in the MERLIN-pedigree files. Three single lymphoblastoid cells were phased per individual.

The resulting single-cell derived haplotypes contained significantly more haplotype blocks consisting of fewer consecutive SNPs when compared to the reference haplotypes (Fig.1A, Supplementary Fig. 2). As a result of these false-positive recombination sites the relative amount of autosomal SNPs in haplotype discordance is approximately 17.72 % (± 5.0 % SD) in the computed single-cell haplotypes (Fig.1E, Supplementary Fig. 3), which makes it difficult to discern the exact

positions of maternal genetic crossovers between grandparental homologous chromosomes on the basis of single-cell derived haplotypes (Supplementary Fig. 2).

The discrepancies between the single-cell derived haplotypes and the reference allelic sequences are due to the genotyping errors of the single-cell MDA-DNA. To derive more reliable genotypes, an algorithm was designed that creates virtual genotypes consisting of SNP-calls that are concordant in two single-cell genotypes of the same individual. On average 96.32 % (± 2.88 % SD) of the homozygous SNPs and 98.46 % (± 1.11 % SD) of the heterozygous SNPs in this dual-cell analysis had a match with the reference genotype (Supplementary Fig.1). Reconstructing haplotypes with these virtual genotypes significantly lowered the number of haplotype blocks, false-positive recombination sites and haplotype discordance of the autosomal SNPs (8.77 % ± 4.24 % SD) (Fig. 1B, 1E, Supplementary Fig. 3), allowing to visually distinguish the true allelic sequence in synteny (Supplementary Fig. 4).

Smoothing of haplotypes

Means to further simplify the analysis of single- as well as dual-cell derived sequences of syntenic alleles were developed. Phasing genotypes of single cells generates numerous small haplotype blocks that do not match the reference haplotype. To evaluate the potential for generating an algorithm that rectifies at least part of these discrepant allelic assignments, histograms were created for the sizes of the haplotype blocks that are respectively concordant, discordant and spanning homologous recombination sites when compared to the reference haplotype (Fig. 2A). This revealed that the majority of haplotype blocks without a reference match are smaller than 200 consecutive SNPs (Fig. 2A-C). Based on this observation, a three-step smoothing algorithm for single-cell haplotypes was designed that re-adjusts the allelic allocation of small haplotype blocks to the grandparental assignments of the flanking large haplotype blocks. In a first cycle of the algorithm, the user-tunable threshold for haplotype block size was set to 200 SNPs. As such, this algorithm will remove potential false positive recombination sites between two flanking haplotype blocks larger than 200 SNPs only if the latter are assigned to the same grandparent. Furthermore, the grandparental origin of the haplotype blocks smaller than the threshold and located at the p- or q-terminus of a chromosome will be adjusted to the first haplotype block at least 200 SNPs wide if the p- or q-terminal haplotype blocks below the threshold are collectively smaller than 500 SNPs. To refine the allelic assignments between

haplotype blocks larger than 200 SNPs and allocated to a different grandparent, thresholds of 50 and 200 SNPs were applied on the smoothed haplotypes in a second and third cycle of the algorithm respectively.

Haplotypes derived from the single- and the dual-cell genotypes were smoothed with this algorithm. Smoothing the syntenic grandparental allelic sequence derived from single-cell genotypes produced less SNPs in haplotype discordance (4.45 % ± 3.85 % SD) (Fig. 1C,E, Supplementary Figs 3, 5) when compared to haplotypes derived from SNP-calls concordant between two single-cell genotypes. Smoothing of the latter dual-cell derived haplotypes produced the best results (SNPs in haplotype discordance: 1.86 % ± 2.13 % SD) (Fig. 1D-E, Supplementary Figs 3, 6).

To determine the accuracy of genetic crossover mapping using the smoothed single- and dual-cell haplotyping data, we aligned 175 homologous recombination sites that were detected in the four individuals. Computation of the frequency of haplotype concordance for 1000 SNPs surrounding these aligned genetic crossovers (Fig. 3A) with the reference allelic sequence, indicated that smoothed dual-cell derived haplotypes pinpoint crossovers more accurately than smoothed single-cell derived haplotypes. Indeed, for smoothed single- and dual-cell derived haplotypes a distance of approximately 250 and 50 SNPs flanking a genetic crossover respectively has to be considered to reach a frequency of >90 % of correct SNP-haplotype inference. The frequency of crossover-flanking SNPs assigned to the correct grandparental origin was higher for the smoothed than the 'raw' single- or dual-cell derived haplotypes (Fig. 3A-B).

In conclusion, this algorithm further facilitates pinpointing homologous recombination sites between grandparental chromosomes on inherited chromatids on the basis of single- or dual-cell analyses.

Haplotyping of human blastomeres

To test whether this single- and dual-cell approach for haplotyping lymphoblastoid cells also work on human blastomeres, the grandparental allelic sequences on the paternally inherited autosomes in three single blastomeres from the same embryo were determined. This male embryo resulted from assisted reproductive technology applied for preimplantation genetic diagnosis (PGD) because the couple wished to circumvent the paternal transmission of a 22q11 deletion to their child (Fig. 4A-B). FISH-analysis on two additional blastomeres revealed that this embryo carried the 22q11 deletion

(data not shown). We were able to confirm this FISH-result by reconstructing the haplotypes of the paternally inherited chromosomes in the three blastomeres.

By probe intensity-cluster graph analysis of SNPs located within the 22q11 deletion it was proven that the deletion occurred *de novo* on a grandmaternal allele in the father (Fig. 4C-F). Haplotype reconstruction of the Dynamic Model-derived genotypes of the three single blastomeres and of three corresponding virtual genotypes resulting from concordant SNP calls in two single blastomeres revealed that the 22q11 locus was embedded in a grandmaternal haplotype block in these blastomeres (Fig. 5A-B). This analysis was performed by including (Supplementary Fig. 7) and excluding (Fig. 5A-B) the SNPs within the known 22q11 deletion region. Since SNPs within the deletion cannot be accurately typed false recombination sites were detected, this was avoided by excluding these SNPs from the haplotype reconstruction. After smoothing of the computed haplotypes, no homologous recombination sites were found on this paternally inherited chromatid of chromosome 22 (Fig. 5C-D). Similar results were obtained for eight other autosomes, while the remaining inherited autosomes showed sites of homologous recombination between grandparental chromosomes (Fig. 6A-D, Supplementary Fig. 7).

Furthermore, a reference haplotype of this embryo was computed from the concordant SNP calls in the three single blastomeres. Approximately 20.3 % (± 3.4 % SD) of the autosomal SNPs in haplotypes derived from single blastomeres were found to be discordant with this reference allelic sequence (Supplementary Fig. 8). Blending the genotypes of these blastomeres two by two lowered the haplotype discordance rate of the autosomal SNPs to 7.54 % (± 1.90 % SD), while smoothing of the single- and dual-cell derived haplotypes yielded discordance rates of 7.84 % (± 1.60 % SD) and 6.08 % (± 1.20 % SD) respectively. Although a reference haplotype derived from non-amplified genomic DNA extracted from multiple cells is lacking for this embryo, these results are comparable to the haplotype reconstructions of the single EBV-transformed lymphoblastoid cells.

In conclusion, our results prove the principle that single- and dual-blastomere haplotyping in combination with haplotype smoothing allow to assess allele-specific inheritance of loci in embryos.

Discussion

We have developed a generic method to genotype a single human cell using high resolution SNP-arrays for subsequent haplotype reconstruction and parental homologous recombination site

identification on the inherited chromatids. As the technique covers the entire genome a full allelic sequence can be determined from one or two cells and thus opens the daunting capability to screen embryos not only for a single Mendelian trait or disease, but also for many Mendelian traits at once as well as for risk factor-, polygenic- or complex disease-variants which are increasingly being uncovered in recent genome-wide association studies.

By smoothing haplotypes that are reconstructed from single-cell derived genotypes the present technique allows the application of linkage-based strategies and to identify parental homologous recombination sites within single lymphoblastoid cells as well as single human blastomeres. In addition, haplotype inference from virtual genomes consisting of SNP-calls that are concordant within two separately genotyped single cells, termed dual-cell haplotyping, with or without subsequent smoothing, adds further reliability to single-cell derived haplotypes enabling the precise demarcation of the majority of the parental homologous recombination sites. We anticipate that increasing the amount of SNP-probes used in these methods together with more stringent SNP-calling thresholds will further improve the phase determination of the single-cell genotype. Furthermore, our data indicate that more efficient genotyping algorithms for single-cell MDA-DNA analyses can be engineered. Apart from single-cell DNA amplification bias, the data shows that a considerable amount of genotyping error is also due to the faulty interpretation of SNP-probe intensities by the genotyping algorithms. We hypothesize that the Dynamic Model outperforms more sophisticated, multi-chip based genotyping algorithms like BRLMM and Birdseed as it calls genotypes by analyzing probe-level intensities one SNP and one array at a time. In this way, single-cell genotype determination is not influenced by the stochastic DNA amplification bias within other single-cell DNA analyses.

The current single- and dual-cell haplotyping methods allow monitoring the inheritance of loci in a generic and genome-wide fashion by typing single-cell MDA-DNA, hence it has the capacity to improve and standardize current preimplantation genetic diagnosis (PGD). PGD is an optional method for couples that wish to circumvent the transmission of heritable diseases to their offspring (Braude et al. 2002). It is underpinned by genetic testing of usually a single or a pair of blastomeres that are biopsied from an IVF embryo on day 3 after fertilization (Renwick and Ogilvie 2007). Embryos without the specific mutation can subsequently be transferred to the woman's uterus on day 4 or 5 after fertilization. Currently, monogenic disorders in pre-implantation embryos are generally diagnosed using PCR amplification of the mutated allele and one or two linked simple tandem repeat (STR)

markers in a multiplex reaction directly on the single-cell DNA (Thornhill et al. 2005) or on multiple displacement amplified (MDA) single-cell DNA (Hellani et al. 2005). However, this approach requires time-consuming designs of family-specific single-cell multiplex PCR reactions that are optimized for minimum allele drop out (ADO), which represents the random failure of amplification of one or multiple alleles at a locus, to avoid misdiagnosis (Rechitsky et al. 1998; Lewis et al. 2001). Furthermore, DNA and amplicon contamination has to be carefully monitored, hence the inclusion of the linked polymorphic markers in the reaction. To get round these difficulties, Renwick et al. (2006) developed preimplantation genetic haplotyping (PGH) which is based on testing family members for multiple polymorphic markers that locate in the close vicinity of a genetic mutation enabling the detection of high-risk and low-risk haplotypes in MDA-amplified single-cell DNA. By including multiple markers in the tests, problems with uninformative markers for specific families, ADO and recombination between a mutation and a linked marker are overcome. However, this technology still requires the development of multiplex PCRs for each locus to be tested. These drawbacks are defeated by our single- and dual-cell haplotyping methods as it uses marker information from thousands of SNPs spread across the genome allowing to follow accurately the segregation of any segment of the genome within a pedigree, thus circumventing labor-intensive designs of family- and locus-specific multiplex PCR reactions. In addition, the technology may not only outperform PCR-based methods to identify the disease allele but also current FISH-based approaches to select embryos suitable for uterine transfer. FISH is applied to select against embryos with unbalanced translocations that result from couples in which one of the parents carries a balanced translocation (Goossens et al. 2004, Melotte et al. 2004). FISH probes have to be developed and validated for every different translocation. In addition, most FISH approaches used for PGD are unable to distinguish those embryos that carry a balanced translocation from those that do not carry the translocation. As such, a number of the embryos suitable for uterine transfer on the basis of FISH still carry the translocated chromosomes. Based on the single/dual-cell haplotyping technology, embryos that carry unbalanced but also the balanced translocations can be identified. Furthermore, for X-linked recessive disorders female embryos are often selected for by FISH (Goossens et al. 2004). However, half of the male embryos disapproved for transfer are free of the disease-causing allele, while half of the selected female embryos still carry the mutation. Knowing the disease causing haplotype enables to sort disease- from non-disease carrying embryos in either sex. Finally, FISH analysis is also used to select against embryos carrying

microdeletions (Goossens et al. 2004; Vanneste et al. 2009b). As shown in this article, the embryos with the microdeletion can easily be identified by the haplotype. In general, all that is required is (1) knowing the risk-haplotype and (2) DNA-samples from relatives to reconstruct the phase of the single- or dual-cell derived genotype. The latter can be achieved by genotyping parental and grandparental DNA samples, but also by SNP-typing parental sperm and polar bodies, DNA from sibs or other family members.

Single- and dual-cell haplotyping in addition broadens the perspective of embryo selection from single Mendelian traits to all phenotypic aspects that are associated with SNP information. Currently, postnatal genome-wide SNP typing is offered to enable disease risk determination for a series of disorders (Kaye 2008). The ability to identify and/or select these traits at the embryonic level raises novel ethical challenges and addressing those challenges becomes pertinent (Klipstein 2005). Although the applicative value of such genomic-phenotypic screens to human embryos is limited, it has huge potential for quantitative trait loci (QTL)-selection in large animal embryos. The majority of important 'performance' traits within livestock and other large animals are underpinned by complex inheritance patterns of multiple quantitative trait loci thus requiring genome-wide marker screens to select the best animal (Goddard and Hayes 2009). Currently, marker-assisted selection uses validated markers in linkage disequilibrium with a Mendelian or quantitative trait locus to select new-born animals. However, due to the limited number of QTLs that are tagged by markers with well-validated associations, its ability to predict breeding values is limited. Recently, a novel method that does not require a complete understanding of complex traits has been developed to predict the genetic merit within an animal by using markers spread over the entire genome that have been typed in a reference population. This genomic selection procedure allows by genome-wide SNP-typing to determine a genomic breeding value that correlates with the true breeding value for a specific animal and is thus expected to have an enormous impact on the speed of breeding programs. Single/dual-cell genotyping and haplotyping methods enable to start this selection process at the embryonic level thus accelerating breeding programs even more by saving unnecessary lengthy periods of gestation.

In addition, this novel single/dual-cell genotyping/haplotyping method allows further characterization of drivers of haplotype diversity, primarily meiotic homologous recombination (Jeffreys et al. 2005; Jeffreys and Neumann 2009), but also mitotic recombination processes that may occur at elevated frequencies during tumorigenesis (Geigl et al. 2008). Inter- as well as intra-chromosomal

rearrangements in somatic cells alter the sequence of syntenic alleles leading to the potential activation of proto-oncogenes and inactivation of tumor suppressor genes. Hence, such recombinations may initiate tumorigenesis, but may also contribute to tumor progression. Due to this chromosome instability tumor cells are heterogeneous and in addition tumor biopsies are contaminated with normal somatic cells. To gain more insight in tumor development and recombination processes single-cell haplotyping is valuable technology.

In conclusion, we developed a generic approach to type single nucleotide polymorphisms and to reconstruct haplotypes on a genome-wide scale using only one or two human cells. This method has the capacity to revolutionize not only genetic selection procedures of both human and animal preimplantation embryos, but also cancer genetic studies and to perform fundamental research on homologous recombination in both human and animal species.

Methods

Collection of lymphoblastoids and blastomeres

Single Epstein-Barr virus (EBV)-transformed lymphoblastoid cells from four individuals were isolated as previously described (Le Caignec et al. 2006). Two unrelated individuals carried a 17q11.2 deletion of 1.3 Mb of maternal origin spanning the *NF1* gene and 12 surrounding genes (Lopez-Correa et al. 2000; De Raedt et al. 2006). The other two individuals carried a 4p16 amplification and had identical parents (Balikova et al. 2008).

The DNA samples of the single human blastomeres have been described in detail previously (Vanneste et al. 2009a). The couple wished to circumvent transmission of a paternal 22q11 deletion to their offspring and entered the IVF-PGD-program. Ovarian stimulation, oocyte retrieval and *in vitro* fertilization procedures were performed as described (Van der Auwera et al. 2002). The embryos were cultured in Global Medium (LifeGlobal). Embryo biopsy of two blastomeres, blastomere fixation and FISH analyses were performed (Vanneste et al. 2009a). We used three remaining blastomeres of a 3-day-old embryo with a microdeletion. The blastomeres were gently disaggregated and treated as described (Le Caignec et al. 2006). The study was approved by the Ethical Committee of the University Hospital Leuven.

Single-cell Affymetrix SNP-array

Single cells were lysed and amplified following a multiple displacement amplification (MDA) approach with Genomi Phi V2 (GE Healthcare) as described (Spits et al. 2006). We did not use single-cell amplifications yielding less than 2 μg DNA.

GeneChip Human Mapping 250K NspI arrays (Affymetrix, GEO accession number GPL3718) were performed in four days according to manufacturer's instructions using 250 ng of single-cell amplified DNA or non-amplified genomic DNA extracted from multiple cells. For SNP-typing, we analyzed obtained probe intensities by (1) the dynamic model algorithm (Di et al. 2005) embedded in the 'GeneChip Genotyping Analysis Software (GTYPE) version 4.1 (Affymetrix)' using a homozygous and heterozygous calling threshold of 0.12, (2) the BRLMM algorithm within the 'Genotyping console 3.0.1' software (Affymetrix) using a scoring threshold of 0.1 and (3) the Birdseed algorithm (Korn et al. 2008) of the APT-1.10.1 package (Affymetrix Power Tools; http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx) using the 'birdseed-dev' command which is the most recently incorporated development version of birdseed from The Broad Institute (http://www.broadinstitute.org/mpg/birdsuite/birdseed.html). The BRLMM and Birdseed clustering algorithms were performed on a batch of 105 MDA-amplified single-cell DNA-samples. BRLMM has a Bayesian step in the Robust Linear Model with Mahalanobis distance classifier (RLMM) (Rabbee and Speed 2006).

Haplotyping

Genome-wide haplotype reconstruction was performed with the MERLIN algorithm using the most likely pattern of gene flow option (**--best** command line option) (Abecasis et al. 2002). The '.ped', '.map' and '.dat' files were generated using Tinn-R 2.2.0.2 and R 2.8.0 (http://cran.r-project.org/). As a requirement of MERLIN (http://www.sph.umich.edu/csg/abecasis/Merlin), every SNP has to have a unique genetic distance. Sex-averaged SNP genetic distances extrapolated from the deCODE map (Kong et al. 2002) were used and are listed in the NetAffx annotation file version NA27 (11/25/08).

Genotype type blending, haplotype smoothing & haplotype alignment

Virtual genotypes consisting of the concordant SNP-calls from two separate single-cell analyses were created using Tinn-R 2.2.0.2 and R 2.8.0. Tinn-R 2.2.0.2 and R 2.8.0 were also applied to align the

allelic assignments of 500 SNPs on either side of 175 homologous recombinations and to compute subsequently the frequency of correct haplotype inference per SNP. The algorithm for smoothing of the single-cell haplotypes was designed in Tinn-R 2.2.0.2 and R 2.8.0. The script is available upon request, for a description of the algorithm see the Results section of this manuscript.

GEO accession number

The SNP-array data has been deposited in NCBI's Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) and is a accessible through GEO Series accession number GSE17318. The blastomere data discussed in this publication is also accessible through GEO Series accession number GSE11663 (Vanneste et al. 2009a).

Acknowledgements

We thank all families who participated in the study, the LUFC for the technical assistance, and S. Jackmaert and M. Amyere for performing the SNP arrays. This work was made possible by grants from the IWT (SBO-60848) and GOA/2006/12 and Center of Excellence SymBioSys (Research Council K.U.Leuven EF/05/007) to J.R.V.. E.V. was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Figure legends

Figure 1. Single- and dual-cell haplotyping, and smoothing of the haplotypes. (A-D) Haplotypes of the maternally inherited chromosome 1 from one individual (indicated with an arrow in the pedigree shown in F). Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. The top lane in each panel represents the reference allelic sequence derived from genomic DNA isolated from multiple cells of this individual. (A) Single lymphoblastoid-cell derived haplotypes of this individual. The three lower lanes show the 'raw' chromosome 1 haplotypes derived from three independent lymphoblastoid cells respectively. (B) Dual lymphoblastoid-cell derived haplotypes. The three lower lanes show the haplotypes derived from three different virtual genotypes respectively. The virtual genotypes consist of concordant SNP-calls of two single lymphoblastoid cells. (C) Smoothed single-cell haplotypes. The three lower lanes show the smoothed chromosome 1 haplotypes derived from three independent lymphoblastoid cells respectively (shown in A). (D) Smoothed results of the haplotypes shown in B. (E) Percentage of autosomal SNPs in haplotype discordance with the reference allelic sequence over the entire genome. Dark blue: single-cell haplotyping, light blue: dual-cell haplotyping, yellow: smoothing of single-cell haplotypes, red: smoothing of dual-cell haplotypes. The depicted values are averages (%) over all 12 lymphoblastoid analyses (± Standard Deviation %). (F) Pedigree structure.

Figure 2. Size distribution of single- and dual-cell derived haplotype blocks. (A) Histograms of single-cell derived haplotype block sizes that are concordant (light green), discordant (red) and spanning a homologous recombination site (pink) when compared to the reference allelic sequence (dark green). The depicted haplotype block sizes are derived from one lymphoblastoid cell analysis. The cutoff of 200 SNPs is indicated in a grey dashed line. (B) Percentage of single- and dual-cell derived haplotype blocks (left and right bars respectively) that are smaller than 200 SNPs and concordant (green), discordant (red) or spanning a homologous recombination site (pink) when compared to the reference haplotype. The depicted values are averages (%) over all 12 single- and dual-lymphoblastoid cell analyses (± Standard Deviation %). (C) Percentage of single- and dual-cell derived haplotype blocks (left and right bars respectively) that are larger than 200 SNPs and concordant (green), discordant (red) or spanning a homologous recombination site (pink) when

compared to the reference haplotype. The depicted values are averages over all 12 single- and dual-lymphoblastoid cell analyses (± Standard Deviation %).

Figure 3. Alignment of single- and dual-cell derived haplotypes at 175 homologous recombination sites. Single- and dual-cell derived haplotypes, both smoothed and nonsmoothed, were aligned at 175 homologous recombination sites identified in the four individuals. Frequencies (%) of crossover-flanking SNPs that are inferred to the correct haplotype according to the reference allelic sequence are depicted. The red line indicates the pooled 175 crossover sites from the reference allelic sequences. SNP haplotype concordance frequencies were computed for 500 SNPs upstream and downstream of the homologous recombination sites. (A) Results obtained with smoothed single- and dual-cell derived haplotypes are depicted in yellow and orange respectively. (B) Results obtained with 'raw' single- and dual-cell derived haplotypes are shown in dark blue and light blue colors respectively.

Figure 4. De novo 22q11 deletion occurred on a grandmaternal allele in the father. (A) Pedigree structure. Grandmaternal and grandpaternal allelic origins are indicated in pink and blue respectively. The arrow indicates the embryo that was diagnosed to carry the 22q11 deletion on the basis of FISH. Three remaining sister blastomeres were evaluated for single- and dual-cell haplotyping. (B) FISH image indicating a 22q11 deletion in a paternal lymphoblastoid nucleus. (C-F) Probe intensity-cluster graph analysis of SNPs located within the 22q11 deletion indicate that the deletion in the father occurred on a grandmaternal chromatid. Color codes are explained in panels C and A.

Figure 5. Single- and dual-blastomere haplotyping identifies the disease-carrying allele in the embryo. (A-D) Haplotypes of the paternally inherited chromosome 22, SNPs within the microdeletion were excluded from the analysis. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. (A) Three single-blastomere derived haplotypes. (B) Three dual-cell haplotypes derived from three different virtual genotypes respectively. The virtual genotypes consist of concordant SNP-calls in two single blastomeres. (C) Smoothing result of the haplotypes shown in B. (D) Smoothing result of the haplotypes shown in D.

Figure 6. Single- and dual-blastomere haplotyping allows pinpointing homologous recombination sites. (A-D) Haplotypes of the paternally inherited chromosome 7. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. (A) Three single-blastomere derived haplotypes. (B) Three dual-cell haplotypes derived from three different virtual genotypes respectively. The virtual genotypes consist of concordant SNP-calls of two single blastomeres. (C) Smoothing result of the haplotypes shown in B. (D) Smoothing result of the haplotypes shown in D.

Supplementary figure legends

Supplementary figure 1. Comparison of genotyping algorithms for single- and dual-cell SNP-typing accuracy. (A-C) Each panel shows for four EBV-transformed lymphoblastoid cell lines the average concordance (± Standard Deviation %) of single-cell homozygous and heterozygous SNPs with the reference genotypes as well as the NoCall rate. The single cells (3 per cell line) were genotyped using either Dynamic Model at threshold 0.12 (A), BRLMM at threshold 0.1 (B) or BIRDSEED at threshold 0.1 (C). (D-F) Similar analysis for dual-cell derived genotypes that consist of SNP-calls concordant in two single-cell genotypes determined using either Dynamic Model at threshold 0.12 (D), BRLMM at threshold 0.1 (E) or BIRDSEED at threshold 0.1 (F). Three different dual-cell derived genotypes were evaluated per cell line.

Supplementary figure 2. Genome-wide single-cell derived haplotypes of four individuals. Per individual the single-cell derived haplotypes of the maternally inherited autosomes and sex chromosome are shown. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. The top lane in each panel represents the reference allelic sequence derived from genomic DNA isolated from multiple cells of a specific individual. The three lower lanes show the 'raw' haplotypes derived from three independent lymphoblastoid cells of a specific individual respectively. Per panel the results for a particular chromosome of a particular individual are shown (individual 1: panels 1-23, individual 2: panels 24-46, individual 3: panels 47-69, individual 4: panels 70-92).

Supplementary figure 3. SNP-haplotype discordance rates in four individuals. Per individual the SNP haplotype discordance rates (± Standard Deviation %) for all maternally inherited chromosomes are shown using either single-cell (n=3) derived haplotypes (dark blue), dual-cell (n=3) derived haplotypes (light blue), smoothed single-cell (n=3) derived haplotypes (yellow) or smoothed dual-cell (n=3) derived haplotypes (red). (A-D) Results for individual 1-4 respectively.

Supplementary figure 4. Genome-wide dual-cell derived haplotypes of four individuals. Per individual the dual-cell derived haplotypes of the maternally inherited autosomes and sex chromosome

are shown. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. The top lane in each panel represents the reference allelic sequence derived from genomic DNA isolated from multiple cells of a specific individual. The three lower lanes show the haplotypes derived from three different virtual genotypes respectively. The virtual genotypes consist of concordant SNP-calls of two single lymphoblastoid cells. Per panel the results for a particular chromosome of a particular individual are shown (individual 1: panels 1-23, individual 2: panels 24-46, individual 3: panels 47-69, individual 4: panels 70-92).

Supplementary figure 5. Genome-wide smoothed single-cell derived haplotypes of four individuals. Per individual the smoothed single-cell derived haplotypes of the maternally inherited autosomes and sex chromosome are shown. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. The top lane in each panel represents the reference allelic sequence derived from genomic DNA isolated from multiple cells of a specific individual. The three lower lanes show the smoothed haplotypes derived from three independent lymphoblastoid cells of a specific individual respectively. Per panel the results for a particular chromosome of a particular individual are shown (individual 1: panels 1-23, individual 2: panels 24-46, individual 3: panels 47-69, individual 4: panels 70-92).

Supplementary figure 6. Genome-wide smoothed dual-cell derived haplotypes of four individuals. Per individual the smoothed dual-cell derived haplotypes of the maternally inherited autosomes and sex chromosome are shown. Alleles that trace back to the grandmaternal and grandpaternal chromosomes are indicated in pink and blue respectively. The top lane in each panel represents the reference allelic sequence derived from genomic DNA isolated from multiple cells of a specific individual. The three lower lanes show the smoothed haplotypes derived from three different virtual genotypes respectively. The virtual genotypes consist of concordant SNP-calls of two single lymphoblastoid cells. Per panel the results for a particular chromosome of a particular individual are shown (individual 1: panels 1-23, individual 2: panels 24-46, individual 3: panels 47-69, individual 4: panels 70-92).

Supplementary figure 7. Genome-wide single- and dual-blastomere haplotyping. Computed haplotypes of the paternally inherited autosomes are shown. Alleles that trace back to the grandmaternal and grandpaternal autosomes are indicated in pink and blue respectively. In consecutive order is shown (one panel per autosome): single-blastomere derived haplotypes (panels 1-22), dual-blastomere derived haplotypes (panels 23-44), smoothed single-blastomere derived haplotypes (panels 45-66) and smoothed dual-blastomere derived haplotypes (panels 67-88).

Supplementary figure 8. SNP-haplotype discordance rates in blastomeres. SNP haplotype discordance rates (± Standard Deviation %) for all paternally inherited autosomes are shown using either single-blastomere (n=3) derived haplotypes (dark blue), dual-blastomere (n=3) derived haplotypes (light blue), smoothed single-blastomere (n=3) derived haplotypes (yellow) or smoothed dual-blastomere (n=3) derived haplotypes (red). A reference haplotype was computed from the concordant SNP-calls among the three single-blastomere derived genotypes.

Table 1. Comparison of genotyping algorithms for single-cell SNP-typing accuracy

Concordance	DM [c]	BRLMM [c]	BIRDSEED [c]
SC-hmz SNPs [a]	89.04 (± 5)	89.50 (± 4.93)	86.97 (± 2.77)
SC-htz SNPs [b]	91.24 (± 6)	67.26 (± 17.29)	48.91 (± 7.32)

[a] Concordance of single-cell homozygous SNPs with the reference genotype
[b] Concordance of single-cell heterozygous SNPs with the reference genotype
[c] Data are represented as mean % (± standard deviation %)
Reference genotypes are Dynamic Model-typed genomic DNA samples (threshold 0.12)

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**: 97-101.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* **322**: 881-888.

Balikova I, Martens K, Melotte C, Amyere M, Van VS, Moreau Y, Vetrie D, Fiegler H, Carter NP, Liehr T, et al. 2008. Autosomal-dominant microtia linked to five tandem copies of a copy-number-variable region at chromosome 4p16. *Am J Hum Genet* **82**: 181-187.

Beaudet AL and Belmont JW. 2008. Array-based DNA diagnostics: let the revolution begin. *Annu Rev Med* **59**: 113-129.

Braude P, Pickering S, Flinter F, Ogilvie CM. 2002. Preimplantation genetic diagnosis. *Nat Rev Genet* **3**: 941-953.

Buntjer JB, Sorensen AP, Peleman JD. 2005. Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci* **10**: 466-471.

Burdick JT, Chen WM, Abecasis GR, Cheung VG. 2006. In silico method for inferring genotypes in pedigrees. *Nat Genet* **38**: 1002-1004.

De Raedt T, Stephens M, Heyns I, Brems H, Thijs D, Messiaen L, Stephens K, Lazaro C, Wimmer K, Kehrer-Sawatzki H, et al. 2006. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nat Genet* **38**: 1419-1423.

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, et al. 2005. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21**: 1958-1963.

Dutt A and Beroukhim R. 2007. Single nucleotide polymorphism array analysis of cancer. *Curr Opin Oncol* **19**: 43-49.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**: 241-251.

Gao G, Allison DB, Hoeschele I. 2009. Haplotyping methods for pedigrees. *Hum Hered* **67**: 248-266.

Garrigan D and Hammer MF. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet* **7**: 669-680.

Geigl JB, Obenauf AC, Schwarzbraun T, Speicher MR. 2008. Defining 'chromosomal instability'. *Trends Genet* **24**: 64-69.

Goddard ME and Hayes BJ. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10**: 381-391.

Goossens V, Harton G, Moutou C, Traeger-Synodinos J, Van Rij M, Harper JC. 2006. ESHRE PGD Consortium data collection IX: cycles from January to December 2006 with pregnancy follow-up to October 2007. *Hum Reprod* **24**: 1786-17810.

Gupta M and Young BD. 2009. Application of SNP genotype arrays to determine somatic changes in cancer. *Methods Mol Biol* **538**: 179-206.

Hellani A, Coskun S, Tbakhi A, Al-Hassan S. 2005. Clinical application of multiple displacement amplification in preimplantation genetic diagnosis. *Reprod Biomed Online* **10**: 376-380.

Iwamoto K, Bundo M, Ueda J, Nakano Y, Ukai W, Hashimoto E, Saito T, Kato T. 2007. Detection of chromosomal structural alterations in single cells by SNP arrays: a systematic survey of amplification bias and optimized workflow. *PLoS.ONE.* **2**: e1306. doi: 10.1371/journal.pone.0001306.

Jeffreys AJ and Neumann R. 2009. The rise and fall of a human recombination hot spot. *Nat Genet* **41**: 625-629.

Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nat Genet* **37**: 601-606.

Jiang Z, Zhang X, Deka R, Jin L. 2005. Genome amplification of single sperm using multiple displacement amplification. *Nucleic Acids Res* **33**: e91. doi: 10.1093/nar/gni089.

Kaye J. 2008. The regulation of direct-to-consumer genetic tests. *Hum Mol Genet* **17**: R180-R183.

Klipstein S. 2005. Preimplantation genetic diagnosis: technological promise and ethical perils. *Fertil Steril* **83**: 1347-1353.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**: 241-247.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**: 1253-1260.

Kotzot D. 2008. Complex and segmental uniparental disomy updated. *J Med Genet* **45**: 545-556.

Le Caignec C, Spits C, Sermon K, DeRycke M., Thienpont B, Debrock S, Staessen C, Moreau Y, Fryns JP, Van SA, et al. 2006. Single-cell chromosomal imbalances detection by array CGH. *Nucleic Acids Res* **34**: e68. doi:10.1093/nar/gkl336.

Lewis CM, Pinel T, Whittaker JC, Handyside AH. 2001. Controlling misdiagnosis errors in preimplantation genetic diagnosis: a comprehensive model encompassing extrinsic and intrinsic sources of error. *Hum Reprod* **16**: 43-50.

Lopez-Correa C, Brems H, Lazaro C, Marynen P, Legius E. 2000. Unequal meiotic crossover: a frequent cause of NF1 microdeletions. *Am J Hum Genet* **66**: 1969-1974.

Mackay I and Powell W. 2007. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* **12**: 57-63.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**: 1166-1174.

Melotte C, Debrock S, D'Hooghe T, Fryns JP, Vermeesch JR. 2004. Preimplantation genetic diagnosis for an insertional translocation carrier. *Hum Reprod* **19**: 2777-2783.

Rabbee N and Speed TP. 2006. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**: 7-12.

Rechitsky S, Strom C, Verlinsky O, Amet T, Ivakhnenko V, Kukharenko V, Kuliev A, Verlinsky Y. 1998. Allele dropout in polar bodies and blastomeres. *J Assist Reprod Genet* **15**: 253-257.

Renwick P and Ogilvie CM. 2007. Preimplantation genetic diagnosis for monogenic diseases: overview and emerging issues. *Expert Rev Mol Diagn* **7**: 33-43.

Renwick PJ, Trussler J, Ostad-Saffari E, Fassihi H, Black C, Braude P, Ogilvie CM, Abbs S. 2006. Proof of principle and first cases using preimplantation genetic haplotyping--a paradigm shift for embryo diagnosis. *Reprod Biomed Online* **13**: 110-119.

Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, Sermon K. 2006. Whole-genome multiple displacement amplification from single cells. *Nat Protoc* **1**: 1965-1970.

Thornhill AR, deDie-Smulders CE, Geraedts JP, Harper JC, Harton GL, Lavery SA, Moutou C, Robinson MD, Schmutzler AG, Scriven PN, et al. 2005. ESHRE PGD Consortium 'Best practice guidelines for clinical preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS)'. *Hum Reprod* **20**: 35-48.

Treff NR, Su J, Tao X, Miller KA, Levy B, Scott RT Jr. 2009. A novel single-cell DNA fingerprinting method successfully distinguishes sibling human embryos. *Ferti. Steril* E-pub ahead of print. doi: 10.1016/j.fertnstert.2009.03.067.

Van der Auwera I, Debrock S, Spiessens C, Afschrift H, Bakelants E, Meuleman C, Meeuwis L, D'Hooghe TM. 2002. A prospective randomized study: day 2 versus day 5 embryo transfer. *Hum Reprod* **17**: 1507-1512.

Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, Melotte C, Debrock S, Amyere M, Vikkula M, Schuit F, et al. 2009a. Chromosome instability is common in human cleavage-stage embryos. *Nat Med* **15**: 577-583.

Vanneste E, Melotte C, Debrock S, D'Hooghe T, Brems H, Fryns JP, Legius E, Vermeesch JR. 2009b. Preimplantation genetic diagnosis using fluorescent in situ hybridization for cancer predisposition syndromes caused by microdeletions. *Hum Reprod* **24**: 1522-1528.

Webb AJ, Berg IL, Jeffreys A. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci USA* **105**: 10471-10476.

Wells D, Alfarawati S, Fragouli E. 2008. Use of comprehensive chromosomal screening for embryo assessment: microarrays and CGH. *Mol Hum Reprod* **14**: 703-710.