



Transcriptome variation in human populations and its potential application in forensics

P. Daca-Roszak¹ · E. Zietkiewicz¹

Received: 14 April 2019 / Revised: 22 July 2019 / Accepted: 24 July 2019 / Published online: 10 August 2019
© The Author(s) 2019

Abstract

This review presents the state-of-the-art in the forensic application of genetic methods driven by the research in population transcriptomics. In the first part of the review, the constraints of using classical genomic markers are shortly reviewed. In the second part, the developments in the field of inter-population diversity at the transcriptomic level are presented. Subsequently, a potential of population-specific transcriptomic markers in forensic science applications, including ascertaining population affiliation of human samples and cell mixtures separation, are presented.

Keywords Transcriptome variation · Human population identification · Laser capture microdissection · Forensic identification

Abbreviations

FID	Forensic identification
AIM	Ancestry Informative marker
LCLs	Lymphoblastoid cell lines
EBV	Epstein-Barr virus
IBD	Markers identical by identical by descent
IBS	Markers identical by state
LCM	Laser capture microdissection

- assessment of the ancestral contribution and of individual's affiliation with continental/ethnic groups
- finding clues about the inherited or acquired phenotype
- identification of the tissue source of a biological material

Genetics in forensic identification

Genetics has been long recognized and adopted as an efficient and reliable approach to forensic identification (FID) of human samples. Genetic-based FID may be perceived from different perspectives, depending on the goal of the investigation. These goals vary considerably and may concern:

- determination of family links
- identification of individuals from whom forensic biological traces derive

Each of the goals in FID requires using appropriate genetic markers and specific methodology, to counteract numerous and variable constraints associated with the analysis (See Fig. 1). Besides the intrinsic constraints associated with the limited information of genetic markers, there are practical concerns in genetic FID, related to any of the following: lack of reference samples, scarcity and/or degradation of the genetic material in forensic traces, and non-homogeneous character of the material (mixed samples).

There is ample literature describing the application of classical, DNA-based genetic markers (microsatellites, SNPs, and haplotypes) for resolving family links (paternity, family relations) and individual's identity as compared with the reference (for the review, see for example Zietkiewicz et al. 2012).

Considerable progress has also been achieved in using DNA markers to assess ancestral contribution to the genome make-up of individuals and thus population affiliation of unknown samples (e.g., Elhaik et al. 2014; Santos et al. 2016). Rapid development of the analysis of human transcriptome and epigenome variability has opened further perspectives, both in the context of tissue identification, and determination of phenotypes; these aspects have been extensively applied in forensic studies (e.g., Frumkin et al. 2011; Xu et al. 2014; Kader and Ghai 2015; Park et al. 2016; Zubakov et al.

Communicated by: Michal Witt

✉ P. Daca-Roszak
patrycja.daca-roszak@igcz.poznan.pl

¹ Institute of Human Genetics, Polish Academy of Sciences, Strzeszynska 32, 60-479 Poznan, Poland

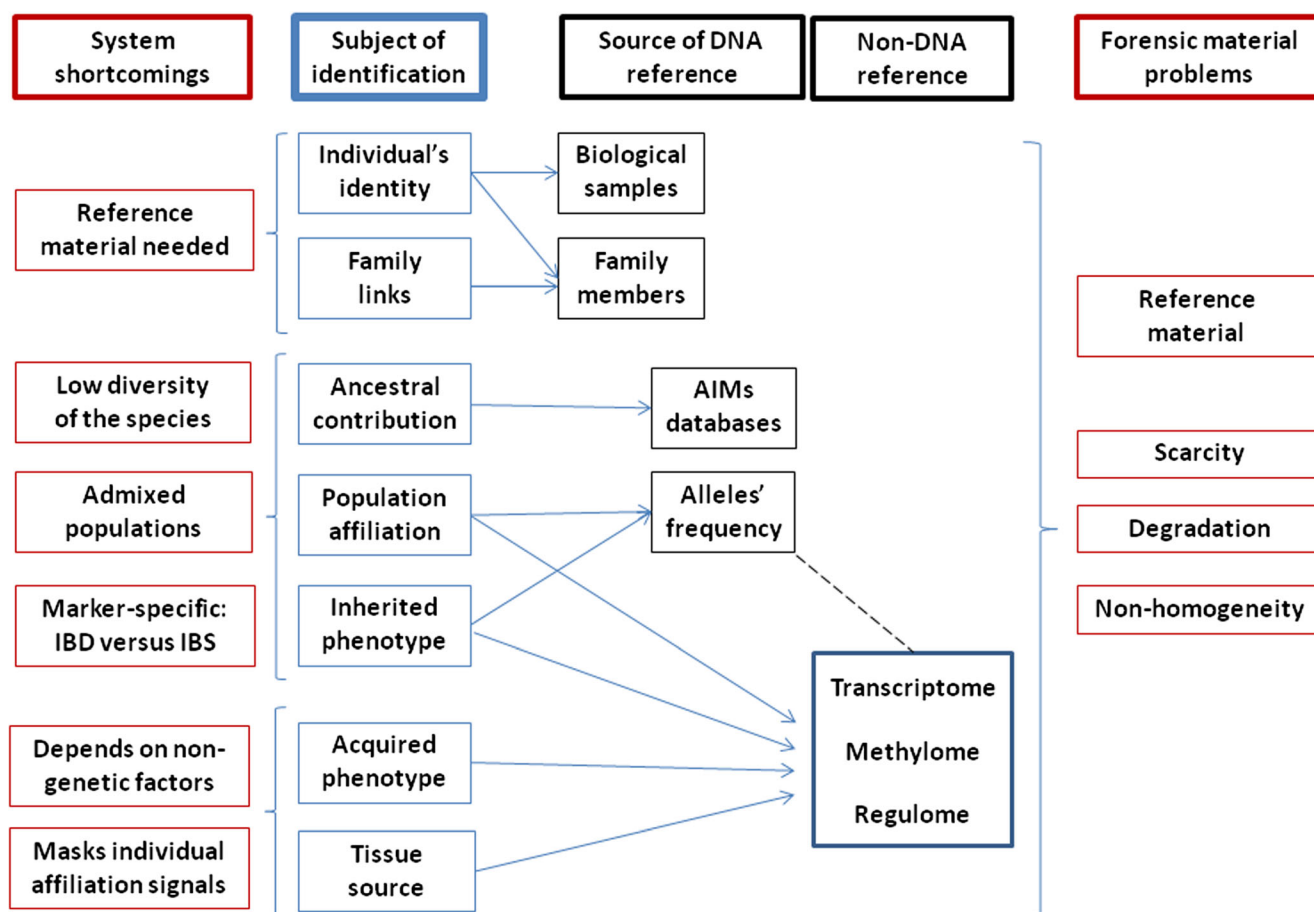


Fig. 1 Forensic identification—schematic representation of the relation between forensic goals and analytical approaches. The different entries in the “subject of identification” column are interdependent—e.g., the

information of the population affiliation or the phenotype of the forensic sample may support the proper identification of the individual

2016). Importantly, both transcriptomic and epigenomics markers may become useful in those of FID endeavors, which require determination of population affinity of the forensic material.

The aim of this review is to present the state-of-the art in the forensic application of genetic methods driven by the research in population genomics and transcriptomics, in the context of some of the major FID problems: the lack of reference samples and non-homogeneity of the biological material.

Genome diversity of human population in FID

The majority of genetic methods used in FID rely on the comparison of a material under investigation with reference samples (from suspected individuals, forensic archives, family members, etc.); sometimes, the information stored in a variety of specialized genetic databases can be used. However, the reference data are often unavailable to an investigator. In such cases, an alternative tactic may be applied: to assign an unidentified biological sample to a specific human population by

comparing it with population-specific data. While indirect, it allows narrowing the focus of the investigation. Consequently, ascertaining population/ethnic affiliation of human samples based on DNA profiling has recently become an important goal in many forensic fields: e.g., crime perpetrator detection, identification of mass disasters or terrorist attack victims (Zietkiewicz et al. 2012; Chakraborty et al. 1999; Budowle et al. 2005; Phillips et al. 2009; Mamedov et al. 2010; Bamshad et al. 2003). The basic shortcoming of population-differentiating genomic markers is related to the low diversity of human species: the majority of genetic variance is shared by all human groups, reflecting the relatively young evolutionary age of our species and/or the recent gene flow (admixture) among extant populations (e.g., Shriver 1997; Zietkiewicz and Labuda 2001; Tishkoff and Williams 2002). In consequence, what actually differentiate human populations are different allele frequencies rather than the presence or absence of marker alleles. Markers with significant frequency differences between human populations are often referred as ancestry informative markers (AIMs) (Frudakis et al. 2003; Shriver and Kittles 2004; Nassir et al. 2009). Relatively few AIMs are

needed for differentiating populations that have diverged a long time ago (e.g., continental groups). However, in case of closely related populations, which share very recent evolutionary history, the sufficient discrimination power can be achieved only by analyzing a very large number of markers distributed over the whole genome (Lao et al. 2006). These analyses usually rely on microarray-based technology (e.g., Novembre et al. 2008; Barbosa et al. 2017).

The majority of population-differentiating markers are selectively neutral, and they mostly reflect the demographic history that shaped the present-day population diversity. AIMS used to infer the ethnic origin of individuals are usually selected from a variety of genomic SNPs, SNP-based haplotypes or CNVs (copy number variants); they may be diploid or haploid (mtDNA, Y-chromosome). Of note, fast-mutating microsatellites (simple tandem repeats (STRs)), which are the most informative markers for the identification of individuals compared with the reference samples, are rarely used as AIMS. This is due to the fact that distinguishing alleles identical by descent (IBD) from those identical by state (IBS) is a challenging task, and conclusions on the population affiliation or the ancestry of the sample are not straightforward.

Many effective population-specific tests have been designed based on markers linked with the genes subjected to selection, e.g., involved in the metabolism of xenobiotics, immune response, fertility, or pigmentation (e.g., Phillips et al. 2007; Rogalla et al. 2015). While these markers can be successfully used to differentiate populations, it has to be remembered that some of the allele frequency similarities, rather than reflecting common ancestry, could be a result of polyphyletic mechanisms that depend on the environment and act in multiple populations independently.

The aforementioned constraints seriously limit the efficiency of DNA-based markers in applications, which require population-based discrimination of the biological material. Recently, intense efforts have been directed to search for non-DNA markers that would exhibit population specificity.

Transcriptome variation among populations

The application of expression microarrays (from Affymetrix or Illumina) targeting thousands of gene transcripts has allowed exploration of the transcriptional variation in humans at the unprecedented scale. First, the levels of gene expression have been shown to differ not only among cells/tissue types, but also among individuals (Cheung et al. 2003; Monks et al. 2004; Morley et al. 2004; Stranger and Dermitzakis 2005). Soon, numerous studies have demonstrated that, while the bulk of variation in the expression level is observed between individuals, significant differences across continental populations also exist (Spielman et al. 2007; Stranger et al. 2007; Storey et al. 2007; Price et al. 2008; Zhang et al. 2008; Ye et al.

2014; Armengol et al. 2009; Fan et al. 2009; Lappalainen et al. 2013; Yin et al. 2014; Mele et al. 2015; Dimas et al. 2009; Li et al. 2014a).

LCL-based studies

The majority of data supporting the notion of the inter-population differences in gene expression are based on the model of lymphoblastoid cell lines (LCLs) (EBV-immortalized human B-lymphocytes). The majority of LCLs, commercially available from Coriell depository and previously used in the International HapMap Project, represent ethnically homogeneous continental populations: CEU—Utah individuals of European ancestry, CHB—Han Chinese, JPT—Japanese, YRI—Nigerians, and AA—admixed African Americans. In spite of the common source of the cell lines used in human transcriptome diversity studies, the direct comparison of the results is difficult for several reasons. First, not all the studies compared the same populations; most often, only limited pairwise comparisons were performed (CEU-CHB; CEU-YRI; YRI-CHB, etc.), and the numbers of individuals representing the populations differed. Second, different methodologies have been applied to determine the expression level (various microarray platforms from Affymetrix and Illumina, or next-generation sequencing (NGS)). Third, estimating and reporting the significance of the results was not uniform (e.g., different statistical models were used; the fold-difference was not always reported; not all the studies provided the names of best-differentiating genes; etc.).

In the seminal study of Spielman, Affymetrix HG-Focus microarray addressing over 4000 genes expressed in lymphoblastoid cells was used to compare expression in Europeans (60 CEU) and Asians (41 CHB and 41 JPT) (Spielman et al. 2007). Over 1000 genes (25%) were found to be differentially expressed between Europeans and Asians (t test, $p < 0.05$), while only 27 genes differentiated Chinese and Japanese. Among 35 genes displaying at least 2-fold expression difference between Europeans and Asians, the best were: *UGT2B17* and *ROBO1* (with 22- and 4-fold higher expression in CEU, respectively) and *CLECSF2* (with 4-fold higher expression in YRI).

In another study using Affymetrix microarray (addressing 5190 genes expressed in LCLs), expression levels in Europeans and Africans (16 CEU and YRI) were compared using models accounting for differences between individuals as well as populations (Storey et al. 2007). Approximately 17% of the genes were differentially expressed in the two populations; the differences in 50 genes were significant at $FDR < 20\%$, with the average fold-change of 1.65. Many of the differentially expressed genes were associated with the immunological response (e.g., gene-encoding cytokines and chemokine receptors: *CCL22*, *CCL5*, *CCR2*, *CXCR3*).

The levels of expression in Europeans and Africans were further analyzed in a larger sample of LCLs (30 CEU and 30 YRI family trios) using Affymetrix Human Exome array (over 9100 transcript clusters) and two independent statistical approaches taking into account the presence of SNPs in the probes (Zhang et al. 2008). About 4.2% of the transcript clusters displayed significantly different expression between the CEU and YRI (247 and 136 with higher expression in YRI and CEU, respectively), with the average fold-change of 1.3. Biological processes found to be enriched in the differential transcripts included ribosome biogenesis and antimicrobial humoral response, as well as cell-cell adhesion, mRNA catabolism, and tRNA processing. Nine of the genes (*DPYSL2*, *CTTN*, *PLCG1*, *SS18*, *SH2B3*, *CPNE9*, *CMAH*, *CXCR3*, and *MRPS7*) were earlier reported among the top 50 genes differentially expressed in CEU and YRI (Storey et al. 2007).

The impact of SNPs and CNV on transcriptome variation has been extensively studied using Illumina whole-genome array in 270 LCLs from CEU, CHB, JPT, and YRI populations (Stranger et al. 2007). Over 5300 genes exceeded the threshold of 16% difference in the median expression in one or more of the population pairs; assuming about 12,000 genes expressed in LCLs, the fraction of genes with significant expression differences between any two populations was estimated between 17 and 29%.

In another study, variation in gene expression was explored in 210 LCLs from four ethnic groups (CEU, CHB, JAP, and AFR), using Illumina microarray addressing more than 11,000 transcripts (Fan et al. 2009). Expression of 427 genes was characterized by higher inter-ethnic than inter-individual variance. Ten of these genes were characterized by the overall variance in expression > 8% (*CXXC4*, *KIF21A*, *LOC376138*, *RGS20*, *TBC1D4*, *TUBB*, *UGT2B11*, *UGT2B17*, *UGT2B7*, and *UTS2*); two of these genes (*UGT2B17*, *RGS20*) have been earlier reported as differentially expressed in Asians and Europeans (Spielman et al. 2007).

After the initial studies based on microarray analysis of transcriptome variability, the dynamic development of high-throughput NGS techniques resulted in a number of studies that analyzed even more transcripts. Besides confirming population differences in the level of expression of a considerable number of genes, they also shed more light upon the mechanisms underlying these differences.

In one of the NGS-based studies (Lappalainen et al. 2013) transcriptomic variation was characterized in over 460 LCLs from Africans (YRI) and four European subpopulations (CEU, FIN, GBR, and TSI). The inter-population differences accounted for only a small fraction (3%) of the total variation in expression. In spite of this, the number of genes displaying significant expression differences between Africans and Europeans was impressively high, ranging from ~1300 to 4300 (depending on which European subpopulation was compared with YRI). The much lower number of differentially

expressed genes was seen when European subpopulations were compared.

In another NGS-based study, expression was examined in 20 LCLs from CEU and CHB (Li et al. 2014a). Over 400 differentially expressed genes were identified (including 132 and 291 with higher and lower expression in CHB, respectively); the magnitude of expression differences was modest (with the median of 2 and 0.4 for the genes up- and downregulated in CHB, respectively). Interestingly, new ethnic-specific isoforms of the known transcripts were revealed in over 200 genes (199 in CHB and 28 in CEU); eleven of those were found in the genes characterized by differential expression in the examined populations (*CLEC2B*, *ARL4C*, *ZBP1*, *ITM2B*, *c11orf21*, *UTS2*, *VCAN*, *CACNA1E*, *EFNA5*, *NR2F2*, and *MGLL*). Ethnic-specific splice junctions were found in only eight genes (*NASP*, *MTIF3*, *CCDC47*, and *TBCA* in CHB and *ITGB7*, *CRTAP*, *ERO1LB*, and *NSUN2* in CEU) (Li et al. 2014a).

In an RNA-sequencing analysis of 45 LCLs from seven non-European populations (Namibian San, Mbuti Pygmies, Algerian Mozabites, Pakistan, East Asia, Siberia, Mexico), 44 differently expressed genes were identified, the vast majority representing immunity pathways. The highest inter-population gene expression variation was obtained for *THNSL2*, *DRP2*, *VAV3*, *IQUB*, *BC038731*, *RAVER2*, *SYT2*, *LOC100129055*, *AK126080*, and *TTN* genes (Martin et al. 2014).

The inter-population differences in the expression level have been repeatedly shown to be heritable and linked to the variation across the human genome. Potential mechanisms include INDELs or copy number variations (CNV) (Spielman et al. 2007; Armengol et al. 2009), SNPs (e.g., Stranger et al. 2007; Storey et al. 2007; Zhang et al. 2008) or alternative splicing (Zhang et al. 2008; Lappalainen et al. 2013; Li et al. 2014a).

Genetic variants in the cis- or trans-acting regulatory elements that affect transcript abundance can be mapped as expression quantitative trait loci (eQTLs). The inter-population variation in these genes' expression are often associated with the population differences in the allele frequencies at eQTLs (Albert and Kruglyak 2015; Kelly et al. 2017; Park et al. 2018). For example, the spectacular difference of *UGT2B17* expression difference between Asians and Europeans was shown to be associated with the higher frequency of the gene deletion in Asians (Spielman et al. 2007). The differential expression of *UGT2B17* locus among populations has also been demonstrated in the study aiming to characterize population differences in the copy number variation (CNV) (Armengol et al. 2009).

Non-LCL studies

All the examples discussed above concerned population differences in gene expression studied in LCLs. In the last few years, several studies have demonstrated that population

differences, similar to those reported in LCLs, are also observed in the cell types other than immortalized B lymphocytes.

In one of these studies, population patterns of gene expression were examined in epidermal samples from 30 individuals representing three continental populations (Yin et al. 2014). Microarray analysis has revealed 14 genes with more than 1.5-fold expression differences between Africans, Caucasians, and Asians. Not surprisingly, the strongest effect was seen between Africans and non-Africans, with 15- and 9-fold differences in the transcription of two best-discriminating genes (*CCL18* and *ADRA2C*, respectively). The differences between Caucasians and Asians were less pronounced, with only one gene, *NINL*, displaying a 3-fold difference in the expression level (Altshuler et al. 2012; Yin et al. 2014).

In another study, focused on population differences in the transcriptional responses of the CD4+ T lymphocytes to the conditions that mimic activation through the antigen-specific receptors (Ye et al. 2014), the set of 236-transcripts was analyzed by Nanostring profiling in 348 donors of African, European, and Asian origin. A trend towards the higher T cell activation in donors of African ancestry has been found to be associated with population differences in the mean expression of a number of genes. For example, expression of *IL2RA* (cytokine receptor) in activated T cells from Africans was approximately 15% higher than in Europeans; other differentially responsive genes included *IL17* family cytokines (over-induced in Africans) and *IFNG* (over-induced in non-Africans).

To exclude the influence of environmental factors (e.g., donor age, time of sampling) on gene expression patterns, the inter-population gene expression variation in placenta was examined in samples from four human populations: African Americans, European-Americans, South Asian Americans, and East Asian Americans (Hughes et al. 2015). The analyses revealed approximately 8% of variation in gene expression among the studied groups. African and South Asian populations had the highest inter-population variation in gene expression (> 140 genes). Genes characterized by the highest inter-population variation were mainly involved in pathways related to immune response, cell signaling, and metabolism (Hughes et al. 2015).

In the recent study on the multi-tissue transcriptomic patterns in Caucasians and Africans, population differences in the expression of over 220 protein-coding genes and 150 lncRNAs (long non-coding RNA) were reported (Mele et al. 2015). However, some of these differentiating markers were specific to individual tissue types. This is consistent with the earlier study, where the direct comparison of gene expression profiles in three types of cells (LCLs, T cells, and primary fibroblasts) has revealed that the majority (80–90%) of genomic variants affecting gene regulation act in a cell type-specific manner (Dimas et al. 2009).

The latter studies have indicated that further surveys are needed to elucidate whether any of the reported population differences in gene expression is common to different cell types. Expression profiling aiming to distinguish ethnic affiliation of the forensic samples would therefore require that the levels of transcripts are compared in the corresponding tissues.

Tissue Expression project (GTEx) may overcome the scarcity of expression data from different human tissues, other than LCLs. GTEx catalogs gene expression variation in major tissues and, additionally, provides an information regarding genetic background underlying this variation (Lonsdale et al. 2013). So far, gene expression profiles for more than 50 human tissues have been cataloged and made publicly available in GTEx database. Hitherto, GTEx project encompasses only data gathered from the Caucasian cohort, which limits applicability of the data to the global population context (Lonsdale et al. 2013).

The application of NGS in human population studies, besides revealing differences in gene expression patterns between distinct human groups (discussed above), provided the knowledge about the diversity of mRNA isoforms in human populations (Park et al. 2018; Djebali et al. 2012; Vaquero-Garcia et al. 2016). It is well known that the vast majority of human genes are subjected to alternative splicing, and a number of isoforms from a single gene may be generated (Pan et al. 2008; Wang et al. 2008; Djebali et al. 2012; Vaquero-Garcia et al. 2016; Park et al. 2018). Various mRNA isoforms have distinct stability and biological function. All these differences in the quantitative and qualitative composition of mRNA isoforms may be adopted as potential population markers.

The latest achievements in the research on alternative splicing variation in human populations have been summarized in the review by Park et al. (2018). The landscape of alternative splicing in relation to the genetic variation has been investigated in a few studies (e.g., Martin et al. 2014; Lappalainen et al. 2013; Battle et al. 2014). Most of the studies were conducted on LCL samples, and they concentrated on the mechanisms underlying formation of RNA isoforms (e.g., Montgomery et al. 2010; Pickrell et al. 2010).

For example, over 170 genes with transcript isoforms changes were identified in the European population (Kwan et al. 2008). Another study has shown that the majority (75 ± 22%) of population-specific variance in gene expression levels observed among seven global human populations can be explained by the variation in gene expression, while only the minor part is caused by the alternative splicing (Martin et al. 2014). This observation was also confirmed in the study of lymphoblastic cells from 69 Yoruban and 60 Caucasian individuals, where RNA sequencing identified 44 genes, for which the ratios of splicing isoforms were similar within each population but more different when comparing populations (Gonzalez-Porta et al. 2012).

The literature data presented above clearly indicate that a significant variation in the expression level across populations exists and that it is at least partially caused by the genomic variation.

The use of specific mRNA transcripts allows efficient differentiation of samples that originate from human populations. Our recent study has revealed two such population-discriminating transcripts: *UTS2* and *UGT2B17* (Daca-Roszak et al. 2018). These mRNA markers exhibited significant population differences in the expression level in both B cell lines and in the peripheral blood and enabled differentiation of Caucasian and Chinese cohorts with high specificity (> 90%) and sensitivity (> 76%) (Daca-Roszak et al. 2018).

Population-specific transcriptome variation—prospects for FID applications

The aforementioned data indicate that carefully chosen population-specific transcriptomic markers can be used in FID applications in a similar way to the DNA-based AIMs, to indicate the population origin of a forensic sample. On the other hand, transcriptomic markers are, just like population-specific DNA markers, more quantitative than qualitative. Moreover, they are expected to be even more susceptible to environmental influences (age, diet). On the other hand, as

discussed below, population-specific transcriptomic markers harbor an important, new potential, which may be prospectively used to solve the great problem of FID application, that of sample mixtures.

The effective use of population-specific genetic markers in FID is often hampered by the non-homogeneity of a forensic material. While deconvolution of allelic profiles obtained from mixed samples is possible (e.g., Hu et al. 2014; van der Gaag et al. 2016), it remains a difficult task and often requires using sophisticated mathematical models (e.g., Bille et al. 2014; Bieber et al. 2016). In practice, identification of multiple contributors by genotyping DNA markers in forensic samples is challenging or not feasible if the reference DNA profiles are not available (Fregeau et al. 2003; Westen et al. 2009). All these features limit the direct use of genetic markers for the analysis of evidentiary samples, which often contain mixed genetic material of unknown origin. Mixtures of cells from the same tissue type, originating from different individuals, are often encountered in the forensic evidence; in the absence of the reference samples, distinguishing the population origin of the individuals, who are the source of such material, poses a serious problem in the FID practice (e.g., Bieber et al. 2016; Gill et al. 2006).

Physical separation of DNA mixtures can be used to address complex DNA mixture problem. In fact various single cell separation technologies have been used before, mainly in

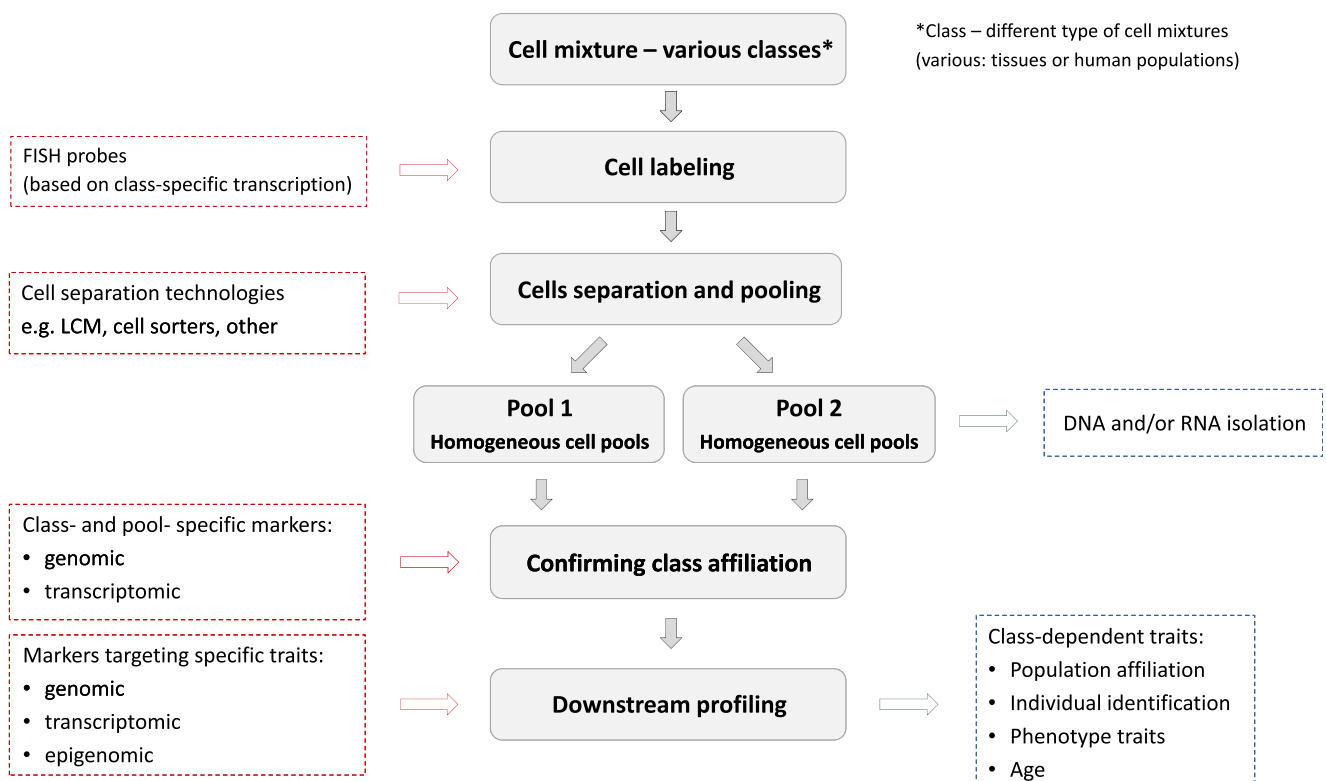


Fig. 2 Potential application of transcriptome variation for mixture deconvolution with the use of LCM technology and FISH labeling

sexual offense cases (e.g., Li et al. 2014b; Williamson et al. 2018) and other examples of tissue separation. However, such idea is brand new in a population-discrimination aspect.

A new perspective is related to a potential application of transcripts characterized by population-specific differences in the expression level. The idea relies on the combination of two techniques: labeling transcripts with population-specific probes and separation of the labeled cells.

The cells from donors of different ethnic background could be “barcoded” with the FISH probes that specifically hybridize to transcripts characterized by differential expression in the relevant populations. In the next step, specifically labeled cells could be separated based on using the cell sorters, laser capture microdissection (LCM) technology, or any other cell separation technique (e.g., Fend et al. 1999; Datta et al. 2015) (see Fig. 2).

Population affiliation of the separated cell pools can be then confirmed by using markers appropriate for the analyzed populations. Markers may be chosen from among transcriptomic probes or genomic eQTLs (SNPs or INDELs) that underlie or associate with the differential expression; population-specific genomic markers not associated with the expression differences (e.g., Daca-Roszak et al. 2016) could be also used for this purpose. The homogenized cell pools can be further used for the downstream profiling using individual-specific or phenotype-specific markers (Vidaki et al. 2013; Zubakov et al. 2010; Zbiec-Piekarska et al. 2015; Koch and Wagner 2011; Bocklandt et al. 2011; Hannum et al. 2013; Weidner et al. 2014).

In most of the forensic cases, DNA/RNA co-isolation from the biological material is possible. The feasibility of combined DNA and RNA profiling of body fluids and contact traces, providing information about both the cell type and sample donor identity, has been reported (Lindenbergh et al. 2012). One can envision that the simultaneous analysis of two transcriptomic markers, one differentiating populations and the other—tissues, combined with any cell separation technique, e.g., LCM technology, could be the way to examine forensic cell mixtures.

Prospects of combining application of distinct markers (transcriptomic, genomic, and epigenetic) for FID purposes, however tempting, are not without flaws. From a technical point of view, the application of transcriptional probes in LCM-based separation of forensic mixtures is, at the present moment, time-consuming and expensive and requires highly qualified and experienced staff. When the amount of a material in the mixed sample is large enough, LCM can be replaced by cell sorters; however, the cell-sorter technology is predictably less suitable for the forensic purposes, where typically only a scarce amount of the evidentiary material is available.

So far, the use of population-specific transcriptomic markers and probes has not been tested in practical forensic applications. The majority of studies were performed in LCLs,

cultured under specific laboratory conditions and derived from a limited set of, mostly continental, populations (e.g., Spielman et al. 2007; Stranger et al. 2007; Storey et al. 2007). Further studies are therefore required to assess sensitivity, specificity, and stability of population-specific transcriptomic markers in real-life samples, which may contain different cell types, like full blood, epithelium, sperm, and hair. Furthermore, additional search for transcripts differentiating more closely related and/or admixed human groups have to be performed. Other problems are related to the non-uniform biological basis of the transcriptome variance (e.g., some transcripts’ levels depend on the environmental factors). Therefore, when selecting transcriptomic markers to be used in the assessment of population affiliation, it will be important to exclude the genes whose expression is known to depend on gender and environmental conditions (diet, stress, etc.).

All the limitation notwithstanding, further exploration of the population-specific transcriptome variation should be the goal of research aiming to improve the applicative prospects in the field of forensic identification.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16(4):197–212. <https://doi.org/10.1038/nrg3891>
- Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65. <https://doi.org/10.1038/nature11632>
- Armengol L, Villatoro S, Gonzalez JR, Pantano L, Garcia-Aragones M, Rabionet R et al (2009) Identification of copy number variants defining genomic differences among major human groups. *PLoS One* 4(9). <https://doi.org/10.1371/journal.pone.0007230>
- Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. *Am J Hum Genet* 72(3):578–589. <https://doi.org/10.1086/368061>
- Barbosa FB, Cagnin NF, Simioni M, Farias AA, Torres FR, Molck MC et al (2017) Ancestry informative marker panel to estimate population stratification using genome-wide human Array. *Ann Hum Genet* 81(6):225–233. <https://doi.org/10.1111/ahg.12208>

- Battle A, Mostafavi S, Zhu XW, Potash JB, Weissman MM, McCormick C et al (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24(1):14–24. <https://doi.org/10.1101/gr.155192.113>
- Bieber FR, Buckleton JS, Budowle B, Butler JM, Coble MD (2016) Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genet* 17. <https://doi.org/10.1186/s12863-016-0429-7>
- Bille TW, Weitz SM, Coble MD, Buckleton J, Bright JA (2014) Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. *Electrophoresis* 35(21–22):3125–3133. <https://doi.org/10.1002/elps.201400110>
- Bocklandt S, Lin W, Sehl ME, Sanchez FJ, Sinsheimer JS, Horvath S et al (2011) Epigenetic predictor of age. *PLoS One* 6(6). <https://doi.org/10.1371/journal.pone.0014821>
- Budowle B, Bieber FR, Eisenberg AJ (2005) Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Leg Med (Tokyo)* 7(4):230–243. <https://doi.org/10.1016/j.legalmed.2005.01.001>
- Chakraborty R, Stivers DN, Su B, Zhong YX, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis* 20(8):1682–1696
- Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M et al (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 33(3):422–425. <https://doi.org/10.1038/ng1094>
- Daca-Rozsak P, Pfeifer A, Zebracka-Gala J, Jarzab B, Witt M, Zietkiewicz E (2016) EurEAS_Gplex-A new SNaPshot assay for continental population discrimination and gender identification. *Forensic Sci Int Genet* 20:89–100. <https://doi.org/10.1016/j.fsigen.2015.10.004>
- Daca-Rozsak P, Swierniak M, Jaksik R, Tyszkiewicz T, Oczko-Wojciechowska M, Zebracka-Gala J et al (2018) Transcriptomic population markers for human population discrimination. *BMC Genet* 19:54. <https://doi.org/10.1186/s12863-018-0663-2>
- Datta S, Malhotra L, Dickerson R, Chaffee S, Sen CK, Roy S (2015) Laser capture microdissection: big data from small samples. *Histol Histopathol* 30(11):1255–1269
- Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H et al (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945):1246–1250. <https://doi.org/10.1126/science.1174148>
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A et al (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108. <https://doi.org/10.1038/nature11233>
- Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Calo CM, De Montis A et al (2014) Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun* 5. <https://doi.org/10.1038/ncomms4513>
- Fan HPY, Di Liao C, Fu BY, Lam LCW, Tang NLS (2009) Interindividual and interethnic variation in genomewide gene expression: insights into the biological variation of gene expression and clinical implications. *Clin Chem* 55(4):774–785. <https://doi.org/10.1373/clinchem.2008.119107>
- Fend F, Emmert-Buck MR, Chuaqui R, Cole K, Lee J, Liotta LA et al (1999) Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. *Am J Pathol* 154(1):61–66. [https://doi.org/10.1016/s0002-9440\(10\)65251-0](https://doi.org/10.1016/s0002-9440(10)65251-0)
- Fregeau CJ, Brown KL, Leclair B, Trudel I, Bishop L, Fourney RM (2003) AmpFI STR (R) profiler PIUS (TM) short tandem repeat DNA analysis of casework samples, mixture samples, and nonhuman DNA samples amplified under reduced PCR volume conditions (25 μ l). *J Forensic Sci* 48(5):1014–1034
- Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S et al (2003) A classifier for the SNP-based inference of ancestry. *J Forensic Sci* 48(4):771–782
- Frumkin D, Wasserstrom A, Budowle B, Davidson A (2011) DNA methylation-based forensic tissue identification. *Forensic Sci Int Genet* 5(5):517–524. <https://doi.org/10.1016/j.fsigen.2010.12.001>
- Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR et al (2006) DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. *Forensic Sci Int* 160(2–3):90–101. <https://doi.org/10.1016/j.forsciint.2006.04.009>
- Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R (2012) Estimation of alternative splicing variability in human populations. *Genome Res* 22(3):528–538. <https://doi.org/10.1101/gr.121947.111>
- Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S et al (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 49(2):359–367. <https://doi.org/10.1016/j.molcel.2012.10.016>
- Hu N, Cong B, Li S, Ma C, Fu L, Zhang X (2014) Current developments in forensic interpretation of mixed DNA samples (review). *Biomed Rep* 2(2049–9434 (Print)):309–316
- Hughes DA, Kircher M, He ZS, Guo S, Fairbrother GL, Moreno CS et al (2015) Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biol* 16. <https://doi.org/10.1186/s13059-015-0627-z>
- Kader F, Ghai M (2015) DNA methylation and application in forensic sciences. *Forensic Sci Int* 249:255–265. <https://doi.org/10.1016/j.forsciint.2015.01.037>
- Kelly DE, Hansen MEB, Tishkoff SA (2017) Global variation in gene expression and the value of diverse sampling. *Curr Opin Syst Biol* 1:102–108. <https://doi.org/10.1016/j.coisb.2016.12.018>
- Koch CM, Wagner W (2011) Epigenetic-aging-signature to determine age in different tissues. *Aging-Us* 3(10):1018–1027
- Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P et al (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* 40(2):225–231. <https://doi.org/10.1038/ng.2007.57>
- Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet* 78(4):680–690. <https://doi.org/10.1086/501531>
- Lappalainen T, Sammeth M, Friedlander MR, t Hoen PAC, Monlong J, Rivas MA et al (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501(7468):506–511. <https://doi.org/10.1038/nature12531>
- Li JW, Lai KP, Ching AKK, Chan TF (2014a) Transcriptome sequencing of Chinese and Caucasian population identifies ethnic-associated differential transcript abundance of heterogeneous nuclear ribonucleoprotein K (hnRNPK). *Genomics* 103(1):56–64. <https://doi.org/10.1016/j.ygeno.2013.12.005>
- Li XB, Wang QS, Feng Y, Ning SH, Miao YY, Wang YQ et al (2014b) Magnetic bead-based separation of sperm from buccal epithelial cells using a monoclonal antibody against MOSPD3. *Int J Legal Med* 128(6):905–911. <https://doi.org/10.1007/s00414-014-0983-3>
- Lindenberg A, de Pagter M, Ramdayal G, Visser M, Zubakov D, Kayser M et al (2012) A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. *Forensic Sci Int Genet* 6(5):565–577. <https://doi.org/10.1016/j.fsigen.2012.01.009>
- Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S et al (2013) The genotype-tissue expression (GTEx) project. *Nat Genet* 45(6):580–585. <https://doi.org/10.1038/ng.2653>
- Mamedov IZ, Shagina IA, Kurnikova MA, Novozhilov SN, Shagin DA, Lebedev YB (2010) A new set of markers for human identification based on 32 polymorphic Alu insertions. *Eur J Hum Genet* 18(7):808–814. <https://doi.org/10.1038/ejhg.2010.22>

- Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee M-C et al (2014) Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genet* 10(8). <https://doi.org/10.1371/journal.pgen.1004549>
- Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M et al (2015) The human transcriptome across tissues and individuals. *Science* 348(6235):660–665. <https://doi.org/10.1126/science.aaa0355>
- Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S et al (2004) Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* 75(6):1094–1105. <https://doi.org/10.1086/426461>
- Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J et al (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289):773–U151. <https://doi.org/10.1038/nature08903>
- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS et al (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001):743–747. <https://doi.org/10.1038/nature02797>
- Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G et al (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet* 10. <https://doi.org/10.1186/1471-2156-10-39>
- Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A et al (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–U5. <https://doi.org/10.1038/nature07331>
- Pan Q, Shai O, Lee LJ, Frey J, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40(12):1413–1415. <https://doi.org/10.1038/ng.259>
- Park J-L, Kim JH, Seo E, Bae DH, Kim S-Y, Lee H-C et al (2016) Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Sci Int Genet* 23:64–70. <https://doi.org/10.1016/j.fsigen.2016.03.005>
- Park E, Pan ZC, Zhang ZJ, Lin L, Xing Y (2018) The expanding landscape of alternative splicing variation in human populations. *Am J Hum Genet* 102(1):11–26. <https://doi.org/10.1016/j.ajhg.2017.11.002>
- Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J et al (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 1(3–4):273–280. <https://doi.org/10.1016/j.fsigen.2007.06.008>
- Phillips C, Prieto L, Fondevila M, Salas A, Gomez-Tato A, Alvarez-Dios J et al (2009) Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS One* 4(8). <https://doi.org/10.1371/journal.pone.0006583>
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E et al (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768–772. <https://doi.org/10.1038/nature08872>
- Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scamici F et al (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* 4(1). <https://doi.org/10.1371/journal.pgen.0030236>
- Rogalla U, Rychlicka E, Derenko MV, Malyarchuk BA, Grzybowski T (2015) Simple and cost-effective 14-loci SNP assay designed for differentiation of European, east Asian and African samples. *Forensic Sci Int Genet* 14:42–49. <https://doi.org/10.1016/j.fsigen.2014.09.009>
- Santos C, Phillips C, Fondevila M, Daniel R, van Oorschot RAH, Burchard EG et al (2016) Paciflex: an ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Sci Int Genet* 20:71–80. <https://doi.org/10.1016/j.fsigen.2015.10.003>
- Shriver MD, Smith MW, Jin J, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, 60:957–964
- Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. *Nat Rev Genet* 5(8):611–6U3. <https://doi.org/10.1038/nrg1405>
- Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet* 39(2):226–231. <https://doi.org/10.1038/ng1955>
- Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM (2007) Gene-expression variation within and among human populations. *Am J Hum Genet* 80(3):502–509. <https://doi.org/10.1086/512017>
- Stranger BE, Dermitzakis ET (2005) The genetics of regulatory variation in the human genome. *Hum Genomics* 2(2):126–131
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813):848–853. <https://doi.org/10.1126/science.1136678>
- Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet* 3(8):611–621. <https://doi.org/10.1038/nrg865>
- van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JFJ et al (2016) Massively parallel sequencing of short tandem repeats-population data and mixture analysis results for the PowerSeq (TM) system. *Forensic Sci Int Genet* 24:86–96. <https://doi.org/10.1016/j.fsigen.2016.05.016>
- Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, Hogenesch JB et al (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 5. <https://doi.org/10.7554/eLife.11752>
- Vidaki A, Daniel B, Court DS (2013) Forensic DNA methylation profiling-potential opportunities and challenges. *Forensic Sci Int Genet* 7(5):499–507. <https://doi.org/10.1016/j.fsigen.2013.05.004>
- Wang ET, Sandberg R, Luo SJ, Khrebtkova I, Zhang L, Mayr C et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476. <https://doi.org/10.1038/nature07509>
- Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P et al (2014) Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol* 15(2). <https://doi.org/10.1186/gb-2014-15-2-r24>
- Westen AA, Matai AS, Laros JFJ, Meiland HC, Jasper M, de Leeuw WJF et al (2009) Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. *Forensic Sci Int Genet* 3(4):233–241. <https://doi.org/10.1016/j.fsigen.2009.02.003>
- Williamson VR, Laris TM, Romano R, Marciano MA (2018) Enhanced DNA mixture deconvolution of sexual offense samples using the DEPArray (TM) system. *Forensic Sci Int Genet* 34:265–276. <https://doi.org/10.1016/j.fsigen.2018.03.001>
- Xu Y, Xie JH, Cao Y, Zhou HG, Ping Y, Chen LK et al (2014) Development of highly sensitive and specific mRNA multiplex system (XCYR1) for forensic human body fluids and tissues identification. *PLoS One* 9(7). <https://doi.org/10.1371/journal.pone.0100123>
- Ye CJ, Feng T, Kwon H-K, Raj T, Wilson M, Asinovski N et al (2014) Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345(6202):1311. <https://doi.org/10.1126/science.1254665>
- Yin L, Coelho SG, Ebsen D, Smuda C, Mahns A, Miller SA et al (2014) Epidermal gene expression and ethnic pigmentation variations among individuals of Asian, European and African ancestry. *Exp Dermatol* 23(10):731–735. <https://doi.org/10.1111/exd.12518>
- Zbiiec-Piekarska R, Spolnicka M, Kupiec T, Makowska Z, Spas A, Parys-Proszek A et al (2015) Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in

- forensic science. *Forensic Sci Int Genet* 14:161–167. <https://doi.org/10.1016/j.fsigen.2014.10.002>
- Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA et al (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet* 82(3):631–640. <https://doi.org/10.1016/j.ajhg.2007.12.015>
- Ziętkiewicz E, Labuda D (2001) Modern human origins in light of the nuclear DNA diversity in world populations. In: Donnelly P, Foley RA (eds) *Genes, fossils and behaviour: an integrated approach to human evolution*. IOS Press, Amsterdam, The Netherlands, pp 79–97
- Ziętkiewicz E, Witt M, Daca P, Zebracka-Gala J, Goniewicz M, Jarzab B et al (2012) Current genetic methodologies in the identification of disaster victims and in forensic analysis. *J Appl Genet* 53(1):41–60. <https://doi.org/10.1007/s13353-011-0068-7>
- Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM et al (2010) Estimating human age from T-cell DNA rearrangements. *Curr Biol* 20(22):R970–R971. <https://doi.org/10.1016/j.cub.2010.10.022>
- Zubakov D, Liu F, Kokmeijer I, Choi Y, van Meurs JBJ, van Ijcken WFJ et al (2016) Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Forensic Sci Int Genet* 24:33–43. <https://doi.org/10.1016/j.fsigen.2016.05.014>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.