16

HPVTyper: A Software Application for Automatic HPV Typing via PCR-RFLP Gel Electrophoresis

Christos Maramis, Dimitrios Karagiannis and Anastasios Delopoulos
Dept. Electrical & Computer Engineering, Aristotle University of Thessaloniki
Greece

1. Introduction

The human papillomavirus (HPV) has been proved to be a many-sided virus. Up to date, over 200 types of HPV have been reported, while ongoing research studies constantly discover new types of the virus. Recently, certain HPV types have been associated with the development of several cancer types (Parkin, 2006), including anal, vaginal, penile, oral, and pharyngeal cancer.

The cervical cancer is currently the most prominent case of association between HPV and cancer. The persistent infection of the anogenital tract by a group of HPV types has been proved to be the main causal factor of cervical cancer (Bosch et al., 2002; Walboomers et al., 1999). This group of cervical cancer related types currently includes 30 to 40 types. However, not all the types in the group induce the same risk for the development of cervical cancer. In fact, virologists have classified the aforementioned types with respect to their oncogenic activity as low-risk, high-risk, and potentially high-risk (Muñoz et al., 2003), while other HPV types remain unclassified.

Focusing on the case of cervical cancer, the statistics are overwhelming: With 500.000 cases diagnosed every year (Greenblatt, 2005), it is the second leading cause of cancer related deaths after breast cancer for women between 20 and 39 years old (Landis et al., 1999) and one of the leading cancer types affecting women worldwide (Moore, 2006). When the above facts are combined with the variability of oncogenic activity among HPV types, the need for specifying not only whether a patient has been infected by HPV but also the exact infecting types becomes evident. The task of identifying the specific HPV type(s) that have infected a patient based on their genotypic characteristics is called *HPV genotyping* or – more simply – *HPV typing*.

In the domain of cervical cancer, HPV typing provides valuable and sometimes critical information to the practitioner of a female patient with respect to her risk for developing cervical cancer. However, it should be noted that as soon as the relation of HPV with other cancer types becomes fully understood, it is highly probable that HPV typing will acquire similar importance in those cancer domains as well. Moreover, HPV typing is also performed for epidemiological purposes, i.e., to determine the frequency and type distribution of the virus in various populations.

The need for HPV typing has given birth to a variety of molecular biology methods and associated assays/kits for carrying out this task. The main objectives that are important when comparing these diagnostic tools are (i) the diagnosis accuracy, (ii) the diagnosis cost, and (iii) the automation level of the diagnostic procedure. Although many solutions are currently available – they will be briefly described in the following section – none stands out when compared to its counterparts. In this chapter, we will present a novel diagnostic tool for HPV typing that attempts to optimize all the aforementioned objectives. This tool is HPVTyper, a freely-available software application that processes images resulting from the PCR-RFLP examination of tissue samples (see Section 2.1) and automatically infers the HPV type(s) that have infected the samples. The novel HPV typing methodology that is employed by HPVTyper has been introduced and extensively validated in a series of recent publications.

2. HPV typing methods

As we have already mentioned, owing to the significance of the task, plenty of molecular diagnostic methods for HPV typing have emerged during the last decades. The majority of them is based on detecting type-specific DNA in an investigated tissue sample (DNA testing methods). The most prominent DNA testing methods are outlined in Section 2.1. The method that is employed by HPVTyper is described separately in Section 2.2.

2.1 Methods overview

Four categories of HPV typing methods are presented in this section. For each category, we provide a short description of the corresponding generic method and comment on its advantages/drawbacks with respect to the other categories.

PCR-based assays. The methods of this category employ the polymerase chain reaction (PCR) to amplify a specific part of the viral DNA. Typing is achieved through the use of type-specific pairs of primers that bound the DNA region to be amplified. Gel electrophoresis of the PCR products is employed to visualize the typing results. The PCR-based methods have been among the first HPV typing methods to be employed and, consequently, many applications of them have been reported (Fontaine et al., 2007; Husnjak et al., 2000; Karlsen et al., 1996; Walboomers et al., 1999).

The main drawback of these methods is the need of a different pair of primers for each type that is considered in the typing process, which, of course, increases the cost of the diagnosis. Moreover, in the case of newly-discovered types appropriate sets of primers have to be designed *de novo*.

Hybridization assays. These methods are based on the hybridization of oligonucleotide probes, where each probe is associated with a specific type of the virus. There are several variants of the hybridization method, namely dot blot (Greer Jr et al., 1990), reverse line blot (Kleter et al., 1999; van den Brule et al., 2002), enzyme immunoassay (Jacobs et al., 1997), etc. These methods employ PCR with general primers in their first stages and their results can be read directly from the assay.

The hybridization methods are less laborious than the PCR-based ones, since they do not require the application of multiple PCRs. However, the previously discussed drawback of the PCR-based methods applies here as well: A different oligonucleotide must be produced – and possibly designed – for each HPV type that needs to be identified.

DNA microarrays. This category of methods shares its conception principle with the hybridization assays (i.e., typing is performed through type-specific oligonucleotide probes) but additionally employs the microarray technology. More specifically, the oligonucleotide probes are "printed" on a two-dimensional array structure and the outcome is usually digitized with the help of a microarray scanner.

The DNA microarray assays have been proved to be very popular during the last decade (Gheit et al., 2006; Hwang et al., 2003; Kim et al., 2003; Klaassen et al., 2004) owing to their accuracy and ease of use. However, being mostly industrial products, the set of HPV types that can be typed by a particular microarray is hardwired into the assay and the expansion to new types necessitates redesigning the assay. Moreover, their consumable nature imposes an additional component to their diagnosis cost.

DNA sequencing. This method discriminates among the HPV types by discovering the exact nucleotide sequence of a small region in the viral DNA. This is the golden standard of virus typing in terms of accuracy since it is able to reveal the viral DNA information to its fullest extent. Since the establishment of the method, several DNA sequencing techniques have been applied on the task of HPV typing (Barzon et al., 2011; Gharizadeh et al., 2001; 2005; Vernon et al., 2000).

The early DNA sequencing techniques have been characterized by low throughput (leading to expensive typing examinations) and could be easily confused in the case of multiple infections. However, the recent introduction of massively parallel sequencing assays (next-generation sequencing) has shown promise to minimize the examination cost and revolutionize DNA sequencing as an HPV typing method (Jordan, 2010; Liu, 2008).

2.2 PCR-RFLP gel electrophoresis

The method of interest for this chapter, i.e., the method that is employed by HPVTyper, is called *PCR-RFLP gel electrophoresis*. The standard protocol for performing HPV typing via the aforementioned method is described in detail in the present section. The protocol involves a series of operations that process the examined tissue sample by means of several established molecular biology techniques and a final step that requires the expertise of a molecular biologist. The involved steps are outlined in Fig. 1(a).

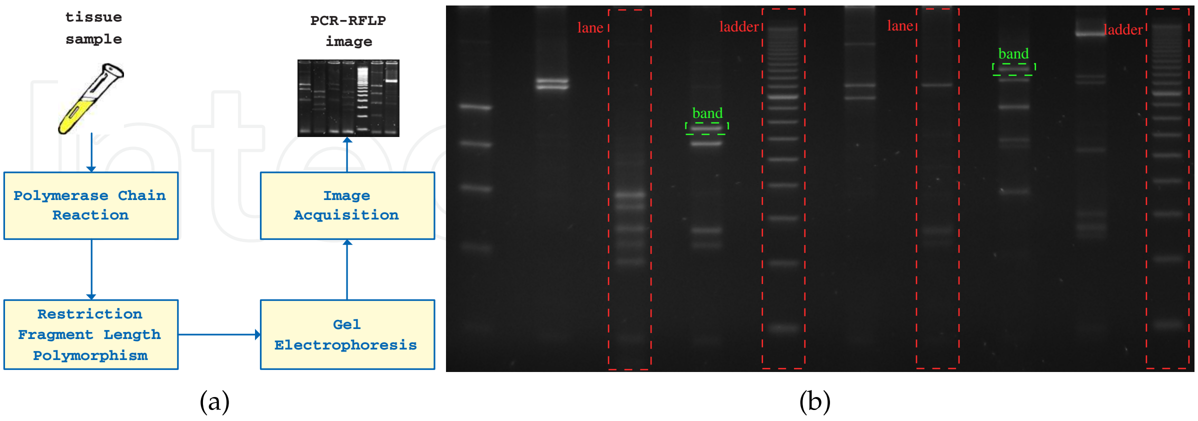


Fig. 1. HPV typing via PCR-RFLP gel electrophoresis. (a) The steps involved in the standard HPV typing protocol. (b) The outcome of a typical PCR-RFLP gel electrophoresis examination; samples of lanes, bands and ladders are enclosed in rectangles.

First, an organic sample is collected from the tissue of interest and the contained DNA is isolated. Then, the PCR (Tagu & Moussard, 2006, Ch. 24) amplifies a highly preserved region of the hypothesized[1] viral DNA with the help of a general-purpose primer set. After that, a restriction enzyme cleaves the amplified HPV DNA at sites that are characterized by a specific nucleotide sequence (restriction sites). This procedure, which is called restriction fragment length polymorphism (RFLP) analysis (Tagu & Moussard, 2006, Ch. 50), produces for each HPV genotype a set of DNA fragments that is known *a priori*; the RFLP analysis is the cornerstone of the discussed method.

Let us elaborate on this. If we assume that the amplified DNA sequence of an HPV genotype of interest is known, then we can predict the sites that will be cleaved by a specific restriction enzyme and, thus, we can infer the set of DNA fragments that are produced after digestion by this restriction enzyme. This set of DNA fragments or *fragment length pattern* (FLP) serves as the signature of this particular HPV genotype and is the means for successfully carrying out the HPV typing.

Following the RFLP analysis, a solution of the digested PCR product is being injected into an individual well at the front end of a gel matrix. Then, in the presence of an electric field, the negatively-charged DNA fragments are forced to move with different mobilities (i.e., drift velocities) towards the anode in the direction opposite to that of the electric field. During the gel electrophoresis (Tagu & Moussard, 2006, Ch. 5), the large molecules remain close to the well, while the more agile smaller molecules cover a much larger distance. This way, one *lane* starting from each well is formed; each lane contains concentrations of DNA of the same length that are shaped as *bands* perpendicular to the electric field direction. One or more among the wells of a gel matrix are reserved to include *DNA ladders*, i.e., DNA molecules of known lengths. These reference ladders assist the biologists in estimating the unknown length of the DNA that forms the bands of the other lanes during the last step of the protocol.

After the electrophoresis is completed, the viral DNA molecules are stained by soaking the gel in a solution of a certain fluorescent dye. For this purpose one can employ either the highly mutagenic yet still common ethidium bromide or, preferably, some other less hazardous fluorescent substance (SYBR Safe, DAPI, etc). Then, an appropriate light source (usually ultra violet or blue) is used momentarily to excite the dye, which fluoresces to make the viral DNA visible. At that moment, a digitized image of the gel matrix is acquired (see Fig.1(b) for an example).

In the final step, the acquired image is analyzed by a molecular biologist in order to reach a typing decision. This analysis includes two stages. First, the biologist estimates the DNA fragment lengths corresponding to the bands that are observed on a lane of interest; this is achieved by interpolating the bands of the image's ladder(s). Then, the biologist "manually" compares the set of estimated fragment lengths from the investigated lane with the FLPs of all the considered HPV genotypes in order to decide which genotype or combination of genotypes has produced the observed band pattern.

From the previous description, one can realize that HPV typing via PCR-RFLP gel electrophoresis suffers from several shortcomings. First of all, it falls short with respect to the other methods in terms of accuracy, since it does not employ type-specific probes or primers. Moreover, it might require a considerable amount of intellectual effort by the molecular biologist to come to a typing conclusion, while typing by the other methods is

[1] We use the word *hypothesized* because the sample can be HPV-free.

essentially straightforward. Finally, the typing procedure becomes even more complicated and error-prone in cases of multiple infections.

Despite the aforementioned shortcomings, PCR-RFLP gel electrophoresis has been used extensively over the past two decades as a method for HPV typing (Lungu et al., 1992; Nobre et al., 2008; Santiago et al., 2006) and remains today the method of choice for a significant percentage of molecular biology laboratories worldwide. This is because the method also demonstrates a series of noteworthy advantages over its counterparts. It does not require the use of overspecialized devices, expensive consumables and type-specific agents. This way it is relatively inexpensive and can be carried out in every moderately equipped molecular biology laboratory. Moreover, owing to its lack of attachment to type-specific entities, the method can easily be exploited for identifying new HPV types, possibly without even the need to repeat the *in vitro* examination of the sample.

3. The employed HPV typing methodology

HPVTyper builds upon a novel HPV typing methodology (Maramis et al., 2010; 2011), which tackles the main shortcomings of typing via the PCR-RFLP method. This methodology is capable of producing very accurate typing decisions – even in complex cases of multiple infections – in an entirely automatic manner. The methodology's performance has been extensively evaluated through a series of experiments (Maramis et al., 2011) on a well-sized set of real HPV data [2].

The discussed methodology is founded on a novel observation model that describes formally the mechanism by which a set of molecular biology parameters (e.g., the concentrations of several HPV genotypes) generates the observed outcome of a PCR-RFLP gel electrophoresis examination (i.e., an image similar to the one depicted in Fig. 1(b)). The observation model is presented in Section 3.1.

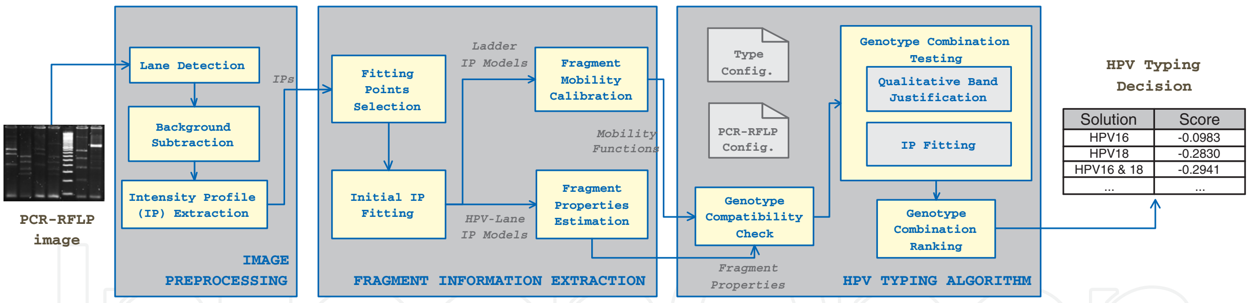
Once the observation model has been established, we proceed with describing the employed methodology. This involves three phases, which are outlined in Fig. 2. In the first phase, the acquired image is preprocessed (Section 3.2). This is a necessary condition for the successful exploitation of the observation model. Then, based on the observation model, accurate information regarding the viral DNA fragments that have resulted from the RFLP analysis is automatically extracted (Section 3.3). The extracted fragment information is fed to a novel HPV typing algorithm (Section 3.4) that identifies – entirely automatically – those HPV genotype(s) that are most probable to have infected the examined sample.

3.1 Observation model

The introduced observation model consists of a set of formulas which associate the visual outcome of a PCR-RFLP gel electrophoresis examination with the underlying parameters of the HPV genotype(s) that have infected the examined sample. In other words, it describes with quantitative terms the mechanism by which the image of such an examination is generated from the infecting HPV genotypes.

In order to present the observation model, let us imagine the following situation: The sample that we are examining has been infected by type T of HPV. After the application of PCR, the

[2] The employed dataset can be accessed via the internet at `http://olympus.ee.auth.gr/~chmaramis/virusTyping`.



Solution	Score
HPV16	-0.0983
HPV18	-0.2830
HPV16 & 18	-0.2941
...	...

Fig. 2. Outline of the employed HPV typing methodology. The main operations are presented as blocks and the operation sequence is indicated by arrows.

concentration of T in the PCR product is c_T. The digestion of the amplified DNA sequence of T by the employed restriction enzyme yields K DNA fragments, whose lengths in base pairs (bp) are contained in the FLP $\boldsymbol{l} = [l_1, l_2, \ldots, l_K]$. At the end of gel electrophoresis, the digested DNA fragments have formed K bands on the sample's lane – this might look like Fig. 3(a) assuming that $K = 4$.

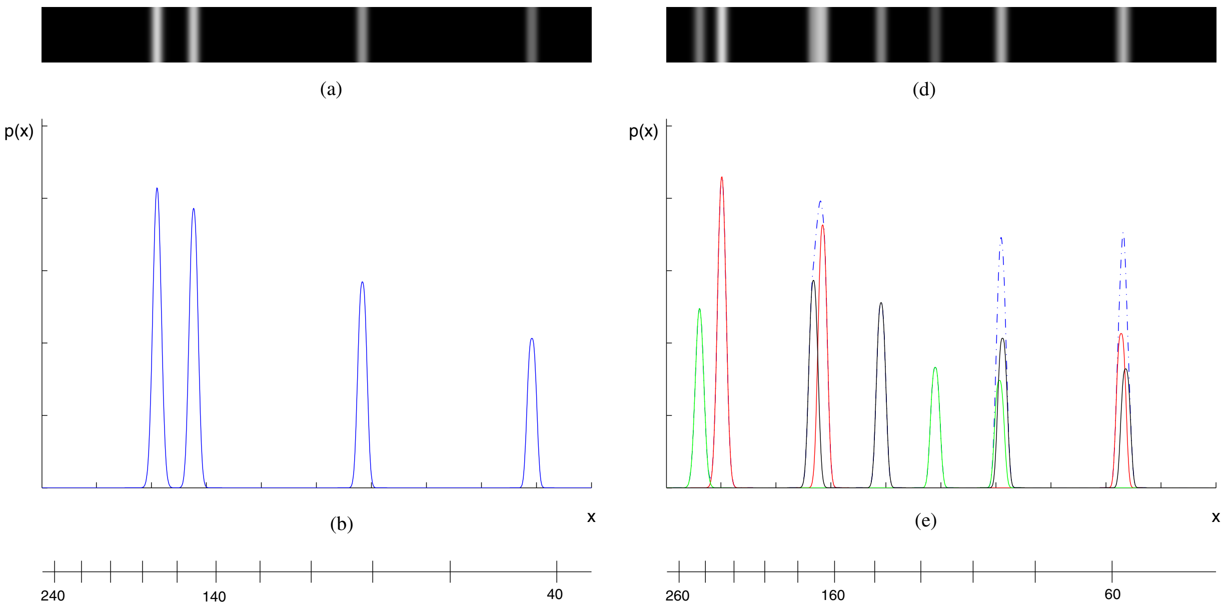


Fig. 3. Illustration of the employed observation model via two examples. (a) Lane image with 4 bands. This corresponds to a sample infected by HPV 53 (GenBank ID X7448) after amplification by MY09/11 (primers) and digestion by HpyCH4V (restriction enzyme). (b) The intensity profile that is extracted from the above lane. The depicted peaks correspond to fragment lengths 171, 151, 83, and 44 bp (left to right). (c) The 20-bp marker that depicts the position-length relation for the above image/profile. (d) Lane image involving triple HPV infection. This corresponds to a sample infected by HPV 30, 62, and 72 (GenBank IDs X74474, AY395706, and X94164) after amplification by MY09/11 and digestion by HpyCH4V. (e) The intensity profile that is extracted from the above lane (blue/dash-dotted line). The contributions of the three infecting types are also displayed with various colors/grayscale levels. (f) The 20-bp marker that depicts the position-length relation for the above image/profile.

Due to the one-dimensional nature of the involved electrophoresis procedure, the information that is conveyed by a lane image lies exclusively in the electrophoresis direction (the X axis in Fig. 3(b)). This means that any two horizontal slices of the lane image are practically replicas of each other. Thus, instead of dealing with the entire lane image, we might as well take a single horizontal slice of it. The resulting one-dimensional intensity curve is called *intensity profile* and is depicted in Fig. 3(b). The intensity profile consists of K bell-shaped peaks, whose centers on axis X are aligned with the positions of the band middles in the lane image.

For a given gel electrophoresis configuration (i.e., fixed gel viscosity, electrophoresis voltage, etc.) the distance that has been covered at the end of the experiment by a certain DNA fragment, and consequently its position, x, on axis X depends only on the length, l, of the fragment. In the described observation model this position-length relation is provided by the following *fragment mobility function*:

$$x = d(l; \boldsymbol{\pi}) = \pi_1 + \pi_2 \log(\pi_3 + \pi_4 l + l^2)\,, \tag{1}$$

where the vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \pi_4]$ aggregates the parameters of the fragment mobility function.

If we assume that $\boldsymbol{\pi}$ is known and recall that the ith peak in the intensity profile of Fig. 3(b) is formed by DNA fragments of length l_i, we conclude that the observed peak centers, $x_1, x_2, \ldots, x_K$, are given by the expression

$$x_i = d(l_i; \boldsymbol{\pi}) \quad i = 1, 2, \ldots, K\,. \tag{2}$$

Moreover, it has been observed that, during gel electrophoresis, a moving population consisting of DNA fragments of some length tends to constantly lose some of its members, i.e., its concentration attenuates continuously as the population moves. This phenomenon is taken into account by the described observation model by means of the residual function

$$r(x) = \phi x + 1 \quad x \in [0, E]\,. \tag{3}$$

This function expresses the percentage of DNA fragments that have not abandoned the moving population when its mean position is x.

In our example we have assumed that the concentration of the PCR product of T is c_T. The PCR product is digested to form K DNA fragment populations that correspond to lengths $l_1, l_2, \ldots, l_K$ and are concentrated in the lane's well before the beginning of the electrophoresis. Thus, the concentrations of these populations at the well will also be:

$$\underbrace{c_T, c_T, \ldots, c_T}_{K}\,.$$

However, due to the aforementioned concentration attenuation, at the end of the electrophoresis, the fragment populations that form the peaks of the intensity profile in Fig. 3(b) will have the following reduced concentrations:

$$\underbrace{r(x_1)c_T, r(x_2)c_T, \ldots, r(x_K)c_T}_{K}\,.$$

The image intensity that is produced by a DNA molecule is determined by the number of fluorescent molecules that are bound to it. Since this number is proportional to the fragment length, the employed observation model computes the contribution of each of the aforementioned K fragment populations to the intensity of the profile as follows:

$$\underbrace{f \cdot l_1 r(x_1) c_T, f \cdot l_2 r(x_2) c_T, \ldots, f \cdot l_K r(x_K) c_T}_{K},$$

where f denotes the contribution of a single-bp DNA fragment to the profile intensity.

Regarding the peak shape, it has been proved (Maramis & Delopoulos, 2010b) that any intensity profile peak corresponding to fragments of length l_0 can be accurately modeled by the integrated Weibull function,

$$w(x;\beta,\gamma,x_0) = \frac{\gamma}{2\gamma^{1/\gamma}\beta\Gamma(1/\gamma)} \exp\Big(-\frac{1}{\gamma}|\frac{x-x_0}{\beta}|^{\gamma}\Big), \tag{4}$$

where (i) $\Gamma(\cdot)$ is the complete gamma function, (ii) the parameters β and γ determine the peak shape, and (iii) x_0 is the mean distance covered by the fragments of length l_0, i.e., $x_0 = d(l_0; \boldsymbol{\pi})$.

Finally, the described observation model superimposes K integrated Weibull functions to model the entire intensity profile of the investigated lane as follows:

$$i(x) = f \cdot c_T \sum_{i=1}^{K} l_i \cdot r(x_i) \cdot w\big(x; \beta_i, \gamma_i, x_i)\big), \tag{5}$$

where $i(\cdot)$ denotes the intensity profile model. In order to derive (5) we have taken into account the previously computed contributions of the DNA fragment populations to the profile intensity.

3.2 Image preprocessing

The first phase of the employed HPV typing methodology deals with the preprocessing of the acquired PCR-RFLP image in order to extract the intensity profiles of all the depicted lanes. The processing operations of this phase presume that the image at hand passes certain quality checks (see Fig. 1(b) for a positive example):

1. The image depicts the lane areas exclusively.
2. The bands appear bright on darker background.
3. The vertical axis of the image identifies with the electrophoresis axis.

Assuming that the image at hand passes these checks – as we will see in Section 4, this might require the application of a few manually-guided image processing operations – the employed methodology detects the boundaries of the lanes that are depicted. Then, with the help of the inter-lane regions it models the background intensity of the image as a polynomial of the image coordinates. Next, it removes the background intensity from the lane areas to keep only the intensity component of the acquired image that is induced by the viral material. All these operations are performed automatically and have been documented in detail elsewhere (Maramis & Delopoulos, 2010a).

This produces a set of background-corrected lane images like the one depicted in Fig. 3(a). The extraction of the intensity profile, $p(\cdot)$, of each lane image concludes this phase. However, due to the presence of noise in the image, instead of taking a single slice of the lane image we aggregate the intensity information along the axis that is perpendicular to electrophoresis (let this be denoted by Y). Thus, if $I(\cdot)$ is the background-corrected image of a lane and D is the image size along axis X, the intensity profile of the lane is given by the expression:

$$p(x) = \underset{y}{\text{median}}\{I(x,y)\} \quad x = 1,2,\ldots,D\,. \tag{6}$$

3.3 Fragment information extraction

The objective of this phase is to estimate certain properties of the DNA fragments that are contained in the lane of a sample. More specifically, we refer to an *initial* estimation of the fragment lengths and concentrations that are associated with the molecule populations on the lane. For this purpose, the employed methodology makes use of the previously extracted intensity profiles and the observation model of Section 3.1. The steps of this phase are described in the following sections.

3.3.1 Peak area detection

This step has not been described in the introductory papers of the employed methodology (Maramis et al., 2010; 2011). It is applied on every intensity profile that has been extracted in Section 3.2 and its objective is to determine (i) the number and *approximate* positions of the peaks in the intensity profile, and (ii) the set of datapoints that constitute the body of each peak. The latter information is utilized by all the subsequent steps of the methodology that involve curve fitting (Sections 3.3.2 and 3.4.2).

First, the rolling disk technique (Mikhailyuk & Razzhivin, 2003) is applied to eliminate from the profile any possible background intensity residuals. The disk radius (parameter θ_1) controls the sensitivity of the aforementioned technique to intensity variations in the profile. After that, the intensity profile is smoothed (θ_2 denotes the order of the smoothing filter) and the watershed algorithm (Vincent & Soille, 1991) detects the local maxima (i.e., peaks) of the intensity profile curve. In order to eliminate possible false peaks (i.e., peaks that do not correspond to real HPV-related bands), a thresholding procedure is employed to rule out peaks lower than a certain multiple (parameter $\theta_3 \geqslant 1$) of the median value of the intensity profile. Next, the watershed algorithm is applied for the second time to detect the pair of the profile's local minima that are adjacent to both sides of each profile peak. Each pair of local minima defines a sequence of profile points which includes the corresponding peak. The central part of this sequence ("representing" the peak's body) will be employed for fitting purposes in the rest of the typing process. The parameter θ_4 specifies the percentage of profile points that are selected for fitting. The selected profile points from all the peaks are aggregated into the set, $\boldsymbol{X}_{\mathrm{F}}$, of the profile's fitting points.

The parameters θ_1 to θ_4 are going to be revisited in the description of HPVTyper in Section 4.

3.3.2 Initial intensity profile fitting

In this step, a simplified model originating from the observation model is employed for each intensity profile that has been extracted – this applies to both the image's ladders and

HPV-related lanes. This model deliberately ignores the – unknown at this point – correlation of the intensity profile peaks through the FLPs of the infecting HPV types. For a lane with K peaks the simplified intensity profile model, $p_m(\cdot)$, is given by the formula

$$p_m(x;\boldsymbol{\rho}) = \sum_{i=1}^{K} A_i \cdot \exp\Big(-\frac{1}{\gamma_i}\big|\frac{x-x_i}{\beta_i}\big|^{\gamma_i}\Big) \quad x = 1,2,\ldots,D\,, \tag{7}$$

where $\boldsymbol{\rho} = [A_1, \beta_1, \gamma_1, x_1, \ldots, A_K, \beta_K, \gamma_K, x_K]$.

The above model is fitted to the extracted intensity profile using the least squares optimization criterion. Formally, this involves finding the parameter vector $\boldsymbol{\rho}_{\mathrm{opt}}$ that satisfies the following equation:

$$\boldsymbol{\rho}_{\mathrm{opt}} = \arg\min_{\boldsymbol{\rho}} \sum_{x \in \boldsymbol{X}_{\mathrm{F}}} \big(p(x) - p_m(x;\boldsymbol{\rho})\big)^2\,. \tag{8}$$

This optimization procedure is called *initial intensity profile fitting* and is carried out once for each lane of the examined image by means of an established optimization algorithm.

3.3.3 Fragment mobility calibration

This step applies only to the ladders of the image and aims to calibrate the fragment mobility function, $d(\cdot)$, of the observation model. In other words, it attempts to calculate the parameter vector $\boldsymbol{\pi}$ (see Section 3.1) that optimally describes the position-length relation of the DNA fragments in the image at hand. Although ideally this relation should be a constant characteristic of the image, the presence of noise and the approximate nature of the fragment mobility function justify the use of all the available ladders to estimate as accurately as possible the position-length relation for various regions of the image.

Focusing on a ladder that includes K peaks, the positions of the peak centers in the intensity profile have been estimated in the previous step; let these be denoted by x_i^* for $i = 1,2,\ldots,K$. Moreover, the ladder specifications provide the molecular lengths that correspond to the aforementioned peaks; let these be expressed as l_i^* for $i = 1,2,\ldots,K$. Then, the optimal parameter vector $\boldsymbol{\pi}_{\mathrm{opt}}$ is given by the expression

$$\boldsymbol{\pi}_{\mathrm{opt}} = \arg\min_{\boldsymbol{\pi}} \sum_{i=1}^{K} \big(x_i^* - d(l_i^*;\boldsymbol{\pi})\big)^2\,. \tag{9}$$

The described *fragment mobility calibration* is performed for each ladder of the image by means of an established optimization algorithm.

3.3.4 Fragment properties estimation

The fragment lengths and concentrations that correspond to the observed populations of HPV DNA molecules (i.e., the observed intensity profile peaks) of each lane are estimated in this step. This applies only to the HPV-related lanes of the experiment. For this purpose, each HPV-related lane is associated with a ladder (possibly the one lying closer to the lane on the gel matrix), which lends the lane its optimized fragment mobility function (Section 3.3.3).

Since the peak positions, x_i, of a profile have been estimated during the initial intensity profile fitting, the fragment lengths, l_i, that are associated with the peaks can be calculated by the inverse of the calibrated fragment mobility function as follows:

$$l_i = d^{(-1)}(x_i; \boldsymbol{\pi}_{\text{opt}}) \quad i = 1, 2, \ldots, K\,, \tag{10}$$

where the parameter vector $\boldsymbol{\pi}_{\text{opt}}$ optimizes the fragment mobility function of the employed ladder and we have assumed that the intensity profile includes K peaks.

Equations (5) and (7) are combined to produce the following formula that estimates the fragment concentrations:

$$c'_i = \frac{2\gamma_i^{1/\gamma_i}\beta_i\Gamma(1/\gamma_i)}{l_i\gamma_i} \cdot A_i \quad i = 1, 2, \ldots, K\,. \tag{11}$$

It is worth mentioning that c'_i denotes the *apparent concentration* of the ith fragment population on the lane, i.e., the concentration that is calculated if we assume that the parameter f in (5) is equal to 1. This assumption does not harm the employed methodology, since we are only interested in the relative concentrations of the observed fragment populations – this will become clear in Section 3.4. For the calculation of the various c'_i, the β_i and γ_i values that have been optimized by (8) and the l_i values that have been estimated by (10) are employed.

3.4 HPV typing algorithm

In this phase the actual typing decisions are made. The *HPV typing algorithm* is charged with the task of deciding which HPV types or combinations of them are able to explain the observed intensity profile peaks of a lane both qualitatively and quantitatively – by means of the observation model. First, all the considered genotypes are checked with respect to their compatibility with the observed peaks (Section 3.4.1) and those that are found incompatible are rejected. Then, each possible combination of compatible genotypes is treated as a hypothesis that is tested for its ability to produce the intensity profile (Section 3.4.2). Finally, the combinations of compatible genotypes are ranked (Section 3.4.3) by a certain score that combines their results in the previous test with the combination's prior probability.

3.4.1 Genotype compatibility check

Let us assume that the intensity profile of the investigated lane has K peaks lying at positions $\boldsymbol{x} = [x_1, x_2, \ldots, x_K]$ and we want to check the compatibility of genotype T with aforementioned profile. If $\boldsymbol{l} = [l_1, l_2, \ldots, l_N]$ denotes the FLP of T, then we define the function $a(\cdot) : [1, \ldots, N] \rightarrow [1, \ldots, K]$ that assigns the FLP components to the profile peaks such that: $a(i) = j$ denotes that the ith component of $\boldsymbol{l}$ *resides* at the jth peak.

Among all possible assignments, we select $a^*(\cdot)$ that minimizes:

$$C_{T,a(\cdot)} = \frac{1}{N}\sum_{i=1}^{N}\left(d(l_i; \boldsymbol{\pi}_{\text{opt}}) - x_{a(i)}\right)^2. \tag{12}$$

Depending on the gel electrophoresis and imaging parameters, the DNA fragments of lengths l_{i_1} and l_{i_2} ($l_{i_1} \neq l_{i_2}$) that correspond to the FLP components with indices $i_1 \neq i_2$ can be perceived as contributing to the same observed peak. However, in the selection of $a^*(\cdot)$ we permit assignment coincidence for components i_1 and i_2 (i.e., $a(i_1) = a(i_2)$) only if the expected positions of fragments i_1 and i_2 are close enough to each other, i.e.,

$$|d(l_{i_1};\boldsymbol{\pi}_{\text{opt}}) - d(l_{i_2};\boldsymbol{\pi}_{\text{opt}})| \leqslant \theta_5 \,. \tag{13}$$

The parameter θ_5 is called *coincidence threshold* and, given a parameter vector $\boldsymbol{\pi}_{\text{opt}}$, determines whether two diverse fragment lengths can be assigned to the same peak or not.

The *compatibility degree* of genotype T is defined as $C_T = C_{T,a^*(\cdot)}$ and has to not exceed the *compatibility threshold* θ_6 should the genotype be considered compatible with the intensity profile (i.e., $C_T \leqslant \theta_6$).

3.4.2 Genotype combination testing

In the combination testing procedure, the phenomenon of *partial digestion* is taken into account. Up to this point, the term FLP has been used to denote the pattern of fragment lengths that results from the digestion of *all* the restriction sites in the amplified DNA molecule (*full digestion*). In this sense, if the amplified molecule of an HPV genotype T contains $N-1$ restriction sites, then the genotype's FLP will be $\boldsymbol{l} = [l_1, l_2, \ldots, l_N]$. However, sometimes restriction enzymes fail to digest the amplified DNA at one or more of the above sites; in these cases FLPs corresponding to partial digestion coexist with the FLP resulting from full digestion. For instance, $\boldsymbol{l}' = [l_1 + l_2, l_3 \ldots, l_N]$ is the result of the enzyme's failure at the first restriction site. To avoid confusion, for the rest of this chapter the terms *main FLP* (mFLP) and *partial digestion FLP* (pdFLP) are employed for full and partial digestion respectively.

For a combination of compatible genotypes $\boldsymbol{t} = \{T_1, \ldots, T_M\}$, $\boldsymbol{l}^k = [l_1^k, \ldots, l_{N_k}^k]$ is employed to denote the mFLP of T_k, with N_k being the number of fragments in $\boldsymbol{l}^k$. With respect to the assignment $a_k^*(\cdot)$ of T_k we adopt the optimal assignment that was selected for this genotype during the compatibility check. Then, the operations that are described in the following paragraphs are performed on combination $\boldsymbol{t}$.

Qualitative peak justification

First, the genotype combination is tested for its ability to qualitatively justify *all* the observed peaks of the profile, with or without the help of a number of pdFLPs that result from the participating HPV genotypes. Only the pdFLPs that result from up to a certain number of cleavage failures (parameter θ_7) are considered. These pdFLPs are combined to form a set of *subhypotheses* so that each subhypothesis includes all the mFLPs of $\boldsymbol{t}$ and a specific subset of the possible pdFLPs. A subhypothesis is accepted at this step if it satisfies the following *qualitative peak justification* criterion:

$$\forall\, j = 1, \ldots, K \quad \exists\, k, i \text{ such that } a_k^*(i) = j \,, \tag{14}$$

where the assignments of the involved pdFLPs (i.e., $a_k^*(\cdot)$ for $k > M$) are also considered; these are selected in the same manner as the mFLP assignments (see Section 3.4.1). Among the subhypotheses that satisfy the above criterion, we select the one that includes the fewest pdFLPs and this is propagated to the following step. If no subhypothesis satisfies the criterion, combination $\boldsymbol{t}$ is rejected.

Intensity profile fitting

In this step, the most crucial of the typing algorithm, the investigated genotype combination hypothesis attempts to explain the intensity profile data as the superposition of contributions from the involved FLPs. In this task, each genotype combination is "represented" by its optimally performing subhypothesis from the previous step.

Let us assume that the selected subhypothesis of the investigated combination $\boldsymbol{t}$ consists of Q FLPs (M mFLPs and $Q - M$ pdFLPs). Based on (5) of the observation model, the individual contribution of the qth FLP $\boldsymbol{l}^q = [l_1^q, \ldots, l_{N_q}^q]$ to the observed intensity profile is given by the equation:

$$i_q(x; c'_q, \boldsymbol{\theta}_q, \phi) = c'_q \sum_{i=1}^{N_q} l_i^q \cdot r(x_i^q) \cdot w(x; \beta_i^q, \gamma_i^q, x_i^q) \quad x = 1, 2, \ldots, D\,, \tag{15}$$

where ϕ is the attenuation factor in (3), c'_q is the apparent concentration of the FLP and $\boldsymbol{\theta}_q = [\beta_1^q, \gamma_1^q, x_1^q, \ldots, \beta_{N_q}^q, \gamma_{N_q}^q, x_{N_q}^q]$.

The intensity profile model of the investigated genotype combination $\boldsymbol{t}$ results from aggregating the contributions of the participating FLPs as follows:

$$\overline{p}_m(x; \overline{\boldsymbol{\rho}}) = \sum_{q=1}^{Q} i_q(x; c'_q, \boldsymbol{\theta}_q, \phi) \quad x = 1, 2, \ldots, D\,, \tag{16}$$

where $\overline{\boldsymbol{\rho}} = [c'_1, \boldsymbol{\theta}_1, \ldots, c'_Q, \boldsymbol{\theta}_Q, \phi]$. The latter vector plays in the present task the same role with the parameter vector $\boldsymbol{\rho}$ of the simplified intensity profile model of (7). The optimal intensity profile model of $\boldsymbol{t}$ results from the following expression:

$$\overline{\boldsymbol{\rho}}_{opt} = \arg\min_{\overline{\boldsymbol{\rho}}} \sum_{x \in \boldsymbol{X}_\mathrm{F}} \left(p(x) - \overline{p}_m(x; \overline{\boldsymbol{\rho}})\right)^2. \tag{17}$$

In the minimization of (17) the following constraint applies: The concentration of a pdFLP is not allowed to exceed a percentage (parameter $\theta_8 < 1$) of the concentration of the corresponding mFLP. The calculation of $\overline{\boldsymbol{\rho}}_\mathrm{opt}$ in (17) under the aforementioned constraint is undertaken by an established optimization algorithm. The initialization of the optimization procedure is performed with the help of the estimates that have been obtained for the fragment properties from the initial intensity profile fitting (Section 3.3.2).

3.4.3 Genotype combination ranking

The score that is employed for ranking the tested combinations of compatible genotypes takes into account both the fitting results of the previous step (Section 3.4.2) and the prior probability of the combination.

Focusing on genotype combination $\boldsymbol{t}$, if $\overline{\boldsymbol{\rho}}_\mathrm{opt}^{\boldsymbol{t}}$ is the optimized parameter vector of the combination's profile model and $\|\boldsymbol{X}_\mathrm{F}\|$ is the number of employed fitting points, the mean squared fitting error of $\boldsymbol{t}$ is defined by the expression

$$J(\boldsymbol{t}) = \frac{1}{\|\boldsymbol{X}_\mathrm{F}\|} \sum_{x \in \boldsymbol{X}_\mathrm{F}} \left(p(x) - \overline{p}_m(x; \overline{\boldsymbol{\rho}}_\mathrm{opt}^{\boldsymbol{t}})\right)^2. \tag{18}$$

Moreover, if J_0 is the mean squared error obtained from the initial fitting of the investigated profile (Section 3.3.2), i.e.,

$$J_0 = \frac{1}{\|\boldsymbol{X}_\mathrm{F}\|} \sum_{x \in \boldsymbol{X}_\mathrm{F}} \left(p(x) - p_m(x; \boldsymbol{\rho}_\mathrm{opt})\right)^2 , \tag{19}$$

$P_0(\boldsymbol{t})$ is the joint prior probability of the HPV types from which the genotypes in $\boldsymbol{t}$ stem, and K is the number of peaks in the profile, then, the score of $\boldsymbol{t}$ is given by the formula

$$R(\boldsymbol{t}) = -2K \frac{J(\boldsymbol{t})}{J_0} + \theta_9 \cdot \ln\left(P_0(\boldsymbol{t})\right) . \tag{20}$$

The parameter θ_9, which is called *prior weight*, is employed to adjust the significance of the prior probability in the ranking process. Its value is positive and defaults to 1.

As one can anticipate from (20), the score only takes negative values, and a genotype combination whose score is close to 0 denotes a better solution than an combination with lower score. The *eligibility threshold* θ_{10} is employed to discriminate between the tested genotype combinations that are considered to be solutions to the problem (i.e., $R(\boldsymbol{t}) \geqslant \theta_{10}$) and those that are not.

4. Description of HPVTyper

HPVTyper is a novel software application for analyzing images that have resulted from the PCR-RFLP examination with the intention of performing accurate and automatic HPV typing. This is achieved by employing the typing methodology that has been presented in Section 3. The application consists of an *engine module* that undertakes all the computations which are involved in the typing methodology (e.g., the initial intensity profile fitting of Section 3.3.2) and a *user interface module* that allows the user to interact graphically with the application. The classic client-server architecture has been employed for implementing the communication between the engine and the user interface.

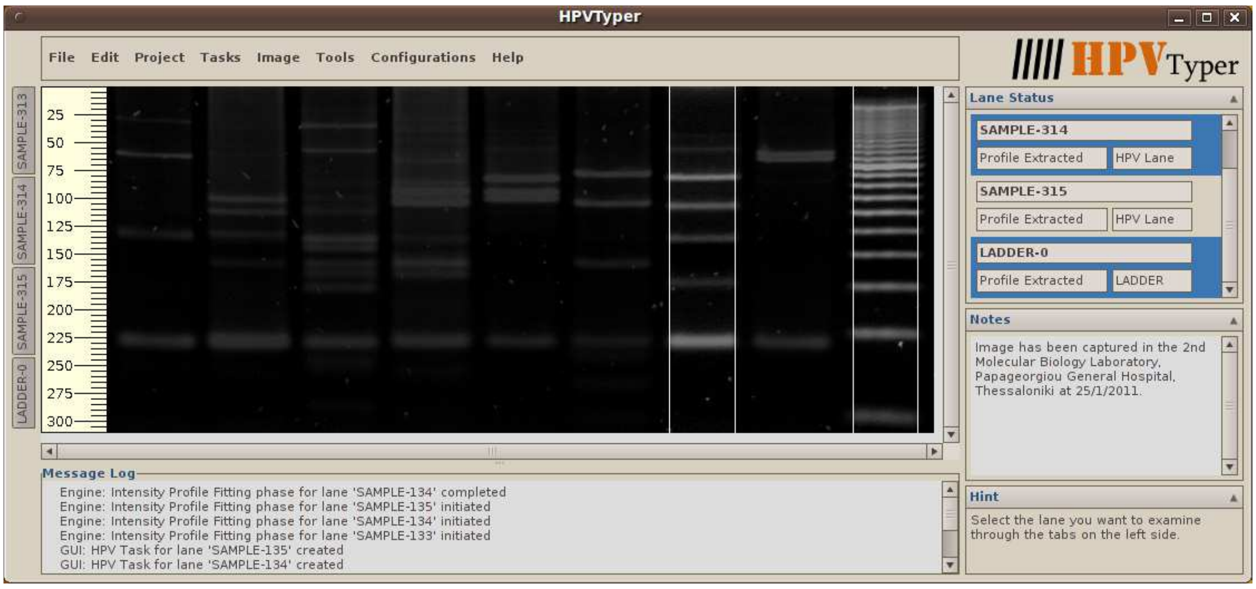
HPVTyper has been designed upon the following principle: The application automates as many of the involved operations as possible while, at the same, it allows the user to intervene essentially at any stage of the typing procedure. This principle is implemented by means of an abundant set of parameters – the already defined parameters θ_1 to θ_{10} are among them. Each parameter possesses a predetermined default value which is used in the involved operations, unless the user wishes to change/adjust it via the graphical user interface.

All the functionalities of the application are made available through four windows, which are described in detail in the following sections.

4.1 Image Processing Window

The main window of HPVTyper is called *Image Processing Window*. This window is presented to the user when HPVTyper initiates and it is mainly concerned with loading the PCR-RFLP images to be analyzed, performing all the required image preprocessing operations on them (Section 3.2), and introducing the depicted lanes. The layout of the Image Processing Window is illustrated in the screenshot of Fig. 4.

As it can be observed in Fig 4, most of the window area is covered by a pane that displays the currently loaded PCR-RFLP image along with a marker indicating positions on the gel



SAMPLE-314 | Profile Extracted | HPV Lane
SAMPLE-315 | Profile Extracted | HPV Lane
LADDER-0 | Profile Extracted | LADDER

Image has been captured in the 2nd Molecular Biology Laboratory, Papageorgiou General Hospital, Thessaloniki at 25/1/2011.

Select the lane you want to examine through the tabs on the left side.

```
Engine: Intensity Profile Fitting phase for lane 'SAMPLE-134' completed
Engine: Intensity Profile Fitting phase for lane 'SAMPLE-135' initiated
Engine: Intensity Profile Fitting phase for lane 'SAMPLE-134' initiated
Engine: Intensity Profile Fitting phase for lane 'SAMPLE-133' initiated
GUI: HPV Task for lane 'SAMPLE-135' created
GUI: HPV Task for lane 'SAMPLE-134' created
```

Fig. 4. Screenshot of the Image Processing Window.

electrophoresis axis (the vertical axis of the image). This is the *image workspace pane* and it can also display certain intermediate image preprocessing results (e.g., the lane boundaries). Below this pane, there is the *message log pane* which displays messages from either the engine module or the user interface module regarding the progress of the HPV typing procedure. With respect to the engine messages, these are the means of the application for notifying the user about changes in the status of computationally expensive operations that are queued in the engine module (e.g., the initiation/termination of the genotype combination testing phase for a particular lane.).

On the right side of the window there is a column that includes three more panes. The first from the top is the *lane status pane*, which allows the user to provide a name for each lane of the active image and to classify it as either HPV-related or ladder. Moreover, it informs the user about the processing status of each lane (e.g., "Boundaries Detected", "Profile Extracted"). When the user selects a lane from the current pane, its position on the loaded image is highlighted as illustrated in Fig. 4. The *notes pane* is located in the middle of the column and allows the user to write down their notes on the present HPV typing assignment in free-text format. The notes are saved and reloaded each time the user opens the same HPV typing project – the concept of HPV typing project will be defined later in the present section. Finally, the *hint pane* contains helpful suggestions of the application to the user regarding the subsequent steps of the typing procedure.

In addition to the aforementioned panes, the Image Processing Window includes a series of menus (ordered in a row on the top of the window) that expose most of the HPVTyper's functionalities to the user.

The most simple functional usage scenario of the main window of HPVTyper is summarized in the following paragraphs. All the described operations are triggered by means of the menu that emerges when right-clicking on the image, unless specified otherwise. First, a PCR-RFLP gel electrophoresis image is loaded to the application through the *File menu*. This action creates an *HPV typing project* dedicated to the typing of the lanes that are included in the image. A number of processing operations can be applied on the image so as to ensure that it passes the quality checks of Section 3.2; these operations include cropping the image, rotating it, and inverting its grayscale levels.

When the image is ready, the user triggers the automatic lane detection operation. Once this operation is completed, the detected lane boundaries are overlaid on the image. At this point, the user can revise/adjust the boundaries of some or all the lanes by moving the displayed boundaries with the help of the mouse. When satisfied with the lane boundaries, the user triggers the background subtraction operation. This is performed automatically and, after its completion, the background-corrected image replaces the original in the image workspace pane.

Next, the user visits the lane status pane in order to provide names for the lanes of interest and classify each of them as HPV-related [3] or ladder. When a lane is selected – by left-clicking on its cell in the lane status pane – HPVTyper displays its position on the image (image workspace pane). By right-clicking on the same cell, the user can ask HPVTyper to extract the intensity profile of the lane. When this happens, HPVTyper creates for this lane either an *HPV typing task*, in case of an HPV-related lane, or a *mobility calibration task*, in case of a ladder. The created task is tied to the selected lane and is uniquely identifiable within the active HPV typing project.

4.2 Profile Processing Window

All the operations that are necessary for carrying out an HPV typing or mobility calibration task are performed in the *Profile Processing Window*. This includes mainly the operations described in Sections 3.3 and 3.4. Each of the HPV typing or mobility calibration tasks that have been created in the main window maintains its own Profile Processing Window. Upon the creation of such a task, a button that activates the corresponding Profile Processing Window is added in a column on the left end of the Image Processing Window.

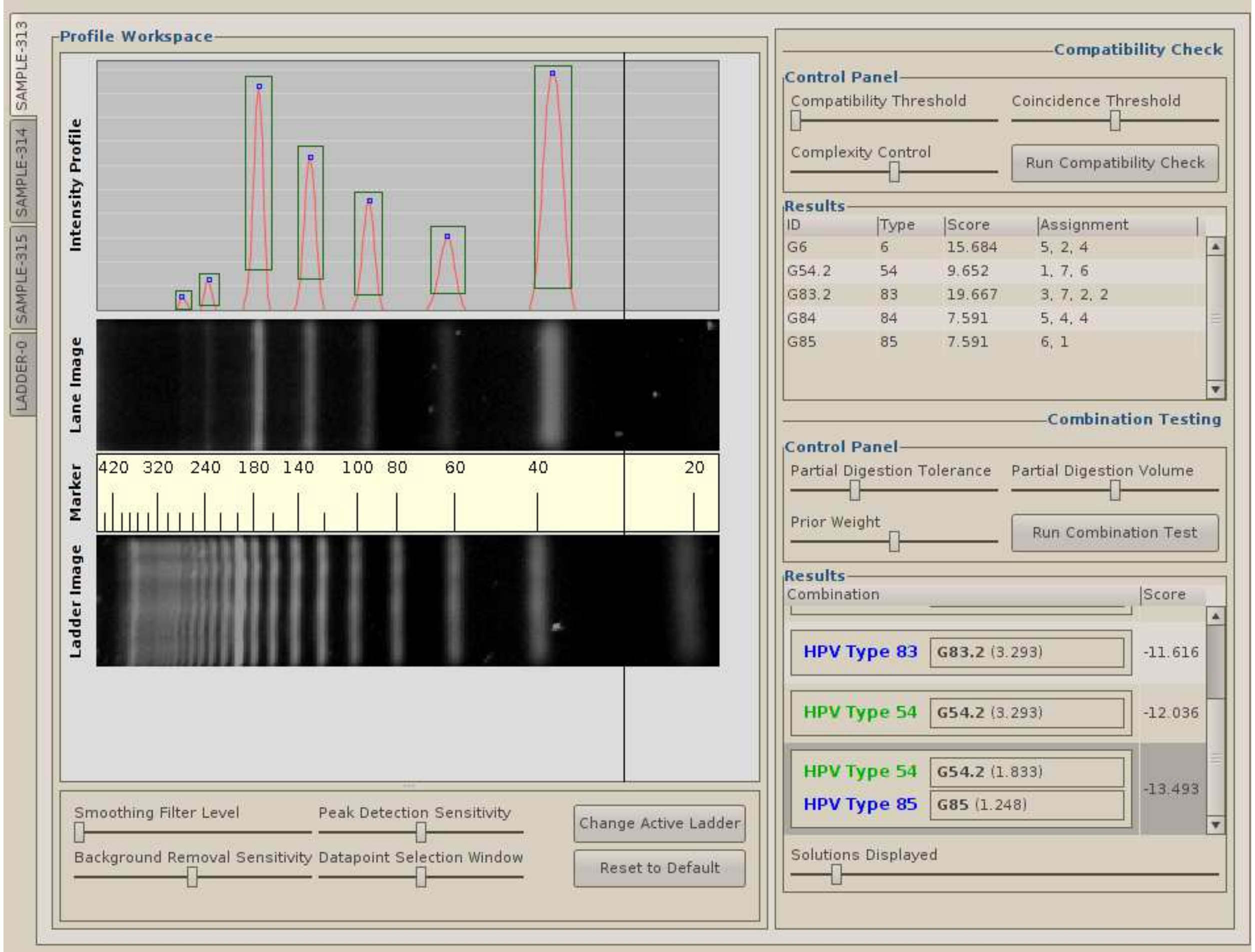
The layout of the Profile Processing Window that corresponds to an HPV-related lane is illustrated in the screenshot of Fig. 5. This includes three panes. Most of the window area is covered by the *profile workspace pane*, which is designed to undertake the fragment information extraction operations described in Section 3.3. On the other hand, the panes in the right column, namely the *compatibility check pane* (top) and the *combination testing pane* (bottom), are meant to carry out the operations corresponding to the HPV typing algorithm (see Section 3.4).

In the profile workspace pane, the horizontal axis identifies with the electrophoresis axis. The pane displays four vertically aligned items, which are (starting from the bottom):

1. The image of the ladder that has been associated with the lane under investigation.
2. The virtual marker that has resulted from the calibration of the aforementioned ladder. This marker specifies the relation between positions and fragment lengths (in bp).
3. The background-corrected image of the investigated lane.
4. The intensity profile that has been extracted from the aforementioned lane.

At the bottom of the profile workspace pane, the user can adjust the values of parameters θ_1 to θ_4 (see Section 3.3.1) and also change the ladder that is associated with the investigated lane.

[3] In the application the term *HPV lane* is employed to denote those lanes that need to be typed.



Results

ID	Type	Score	Assignment
G6	6	15.684	5, 2, 4
G54.2	54	9.652	1, 7, 6
G83.2	83	19.667	3, 7, 2, 2
G84	84	7.591	5, 4, 4
G85	85	7.591	6, 1

Results

Combination		Score
HPV Type 83	G83.2 (3.293)	-11.616
HPV Type 54	G54.2 (3.293)	-12.036
HPV Type 54 HPV Type 85	G54.2 (1.833) G85 (1.248)	-13.493

Fig. 5. Screenshot of the Profile Processing Window.

The compatibility check pane, which deals with the procedure described in Section 3.4.1, is divided into two parts. The upper part is employed to control the compatibility check procedure (parameters θ_5 and θ_6) and determine the maximum number of compatible genotypes that can be considered by the application (*complexity control* slider). This way, the computational complexity involved in the combination testing phase of the HPV typing algorithm can be controlled by the user. In the lower part, the compatibility check results are presented in a table which includes (i) the genotypes that are found to be compatible with the profile, (ii) the HPV types to which they belong, (iii) the scores that they have achieved in the compatibility check, and (iv) the selected assignment, $a^*(\cdot)$, of each genotype's FLP.

The combination testing pane, which serves the operations described in Sections 3.4.2 and 3.4.3, resembles in terms of design the previous pane: It is divided into an upper part that controls the combination testing and ranking procedure and a lower part that displays the results. In the upper part the values of parameters θ_7 to θ_{10} can be defined. In the lower part, the solutions of the HPV typing algorithm are displayed – sorted by their scores – in a table that includes for each solution (i) its score, (ii) the participating genotypes and associated concentrations that achieve this score, and (iii) the HPV types to which the genotypes belong. A color-coding scheme is employed for the types based on their oncogenic risk (e.g., green denotes low-risk types). The maximum number of solutions that can be displayed is specified with the help of a slider at the bottom of the pane.

When it comes to ladders, only a reduced set of operations needs to be supported (Sections 3.3.1 to 3.3.3). For this reason, the Profile Processing Window that is associated with a ladder is slightly different from what has been already described. More specifically, the compatibility check and combination testing panes are disabled and the lane image item is missing from the profile workspace pane. The fragment mobility calibration functionality is provided through the corresponding option in the right-click menu of the aforementioned pane.

The shortest sequence of actions that completes an HPV typing task is briefly presented below. First, the user has to calibrate the mobility function from the ladder that is associated with the HPV-related lane under investigation. This procedure is performed on the ladder's Profile Processing Window. At this point, all the operations described in Section 3.3.1 can be performed automatically by HPVTyper via the right-click menu of the profile workspace pane and afterwards modified manually by the user in a context-aware manner. Once the calibration is achieved, the user switches to the Profile Processing Window of the investigated HPV-related lane. Operations similar to those performed for the ladder can be attempted on this profile as well, before requesting HPVTyper to execute the initial intensity profile fitting process. Following the profile fitting, the application automatically estimates the fragment properties associated with the profile peaks. Then, the compatibility check is triggered by the user and the obtained results are displayed in the appropriate table. Finally, the user has to initiate the combination testing procedure. This is executed by the engine module, and the computed solutions are displayed along with their characteristics in the respective table upon completion of the procedure. The optimal intensity profile model corresponding to each solution can be drawn in the profile workspace pane upon user request, allowing them to visually supervise the end fitting results.

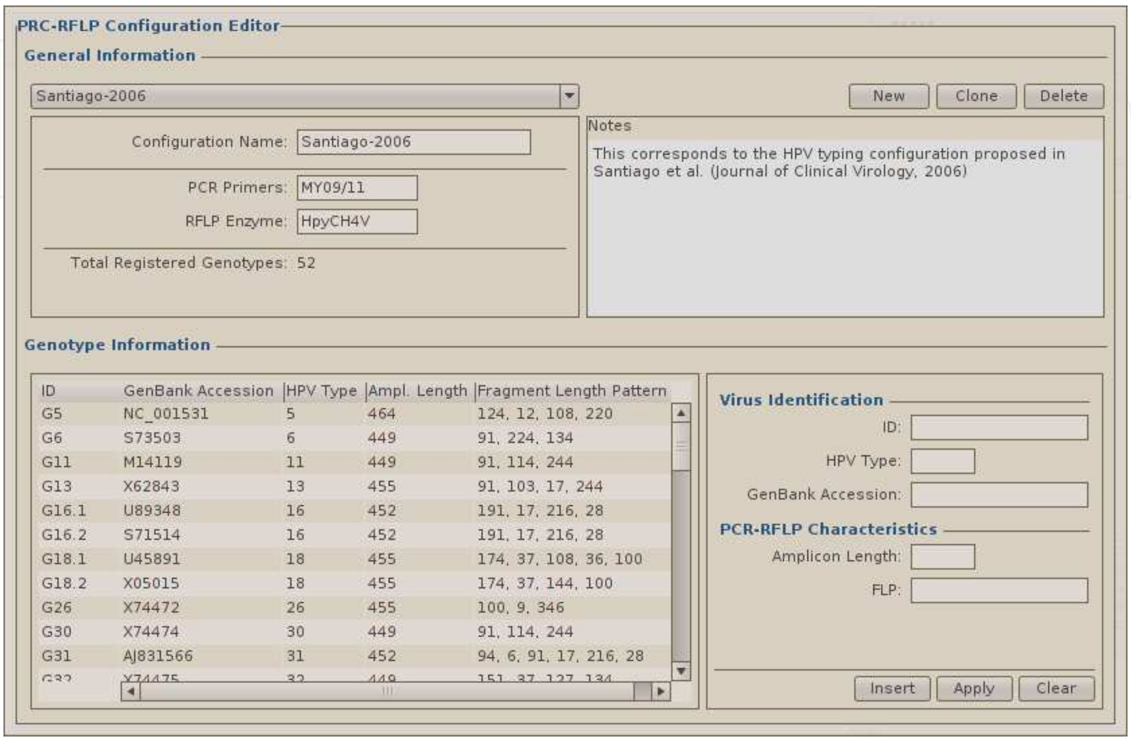
4.3 Configuration Editors

By the term *Configuration Editors* we refer to two distinct windows that are employed for editing the information required for making the HPV typing decisions. These windows, namely the *PCR-RFLP Configuration Editor* and the *Type Configuration Editor*, are meant to be used rarely, e.g., whenever new scientific discoveries need to be taken into account in the typing process. They can both be accessed through the *Configurations menu* of the main window and are described in the following sections.

4.3.1 PCR-RFLP Configuration Editor

Since various combinations of primers and restriction enzymes can be employed for a PCR-RFLP examination and also various HPV genotypes may be considered in an HPV typing task, the knowledge of the above parameters – the *PCR-RFLP configuration* – is necessary when one attempts to perform HPV typing on an acquired PCR-RFLP image. The present window allows the user to insert information regarding the above configuration into HPVTyper. The application can store many different PCR-RFLP configurations and the user is responsible for assigning the appropriate one to each HPV typing task through the *Tasks menu* of the application's main window.

The layout of the PCR-RFLP Configuration Editor is illustrated in the screenshot of Fig. 6 and its usage is straightforward. In the upper part of the window, the user can switch between the stored configurations, keep configuration-related notes, and define the associated PCR primers and restriction enzyme(s). In the lower part of the window, the application displays in a tabular form (i) the HPV genotypes that are considered in the present configuration (user-defined names are employed), (ii) their GenBank accession ID, (iii) the HPV types to which they belong, (iv) the length of their amplicon, and (v) their fragment length pattern. On the right side of this table, there is a form for editing the genotype-specific information.



Santiago-2006

Configuration Name: Santiago-2006

PCR Primers: MY09/11

RFLP Enzyme: HpyCH4V

Total Registered Genotypes: 52

Notes

This corresponds to the HPV typing configuration proposed in Santiago et al. (Journal of Clinical Virology, 2006)

ID	GenBank Accession	HPV Type	Ampl. Length	Fragment Length Pattern
G5	NC_001531	5	464	124, 12, 108, 220
G6	S73503	6	449	91, 224, 134
G11	M14119	11	449	91, 114, 244
G13	X62843	13	455	91, 103, 17, 244
G16.1	U89348	16	452	191, 17, 216, 28
G16.2	S71514	16	452	191, 17, 216, 28
G18.1	U45891	18	455	174, 37, 108, 36, 100
G18.2	X05015	18	455	174, 37, 144, 100
G26	X74472	26	455	100, 9, 346
G30	X74474	30	449	91, 114, 244
G31	AJ831566	31	452	94, 6, 91, 17, 216, 28
G32	X74475	32	449	151, 37, 127, 134

ID:

HPV Type:

GenBank Accession:

Amplicon Length:

FLP:

Fig. 6. Screenshot of the PCR-RFLP Configuration Editor.

4.3.2 Type Configuration Editor

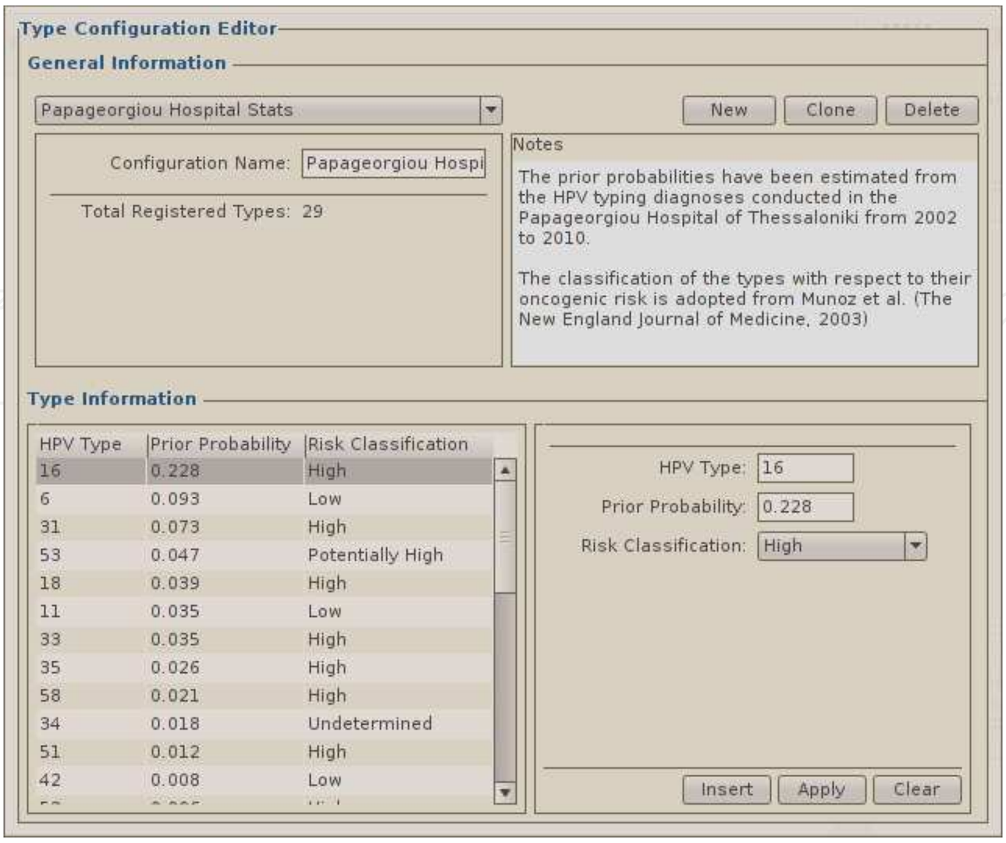
The HPV types demonstrate significant variations in their prevalence in different populations and tissue types. Since the typing methodology that is employed by HPVTyper explicitly uses the prior probabilities of the various HPV types to make its typing decisions, it is reasonable to allow multiple sets of prior probabilities to be included in the application. This way, the appropriate set of prior probabilities that best suits an examined tissue sample by geographic and/or anatomic criteria can be employed.

The present window allows the user to edit an existing set of prior probabilities or insert a new set. In addition, it provides the user with the opportunity to define the oncogenic risk of each HPV type; this information is used in the color-coded representation of the types in the Profile Processing Window.

The design of the present window resembles that of the PCR-RFLP Configuration Editor and all the provided functionalities are straightforward. The layout of the Type Configuration Editor is illustrated in the screenshot of Fig. 7.

5. Important features of HPVTyper

As it may have become clear from the presentation of the employed methodology but also from the description of the application itself, HPVTyper demonstrates several noticeable



Papageorgiou Hospital Stats

Configuration Name: Papageorgiou Hospi

Total Registered Types: 29

Notes

The prior probabilities have been estimated from the HPV typing diagnoses conducted in the Papageorgiou Hospital of Thessaloniki from 2002 to 2010.

The classification of the types with respect to their oncogenic risk is adopted from Munoz et al. (The New England Journal of Medicine, 2003)

HPV Type	Prior Probability	Risk Classification
16	0.228	High
6	0.093	Low
31	0.073	High
53	0.047	Potentially High
18	0.039	High
11	0.035	Low
33	0.035	High
35	0.026	High
58	0.021	High
34	0.018	Undetermined
51	0.012	High
42	0.008	Low

HPV Type: 16

Prior Probability: 0.228

Risk Classification: High

Fig. 7. Screenshot of the Type Configuration Editor.

features. When combined, these features establish HPVTyper as a noteworthy HPV typing approach that is significantly differentiated from its counterparts. The most important of the application's features are outlined in this section.

Accurate typing decisions. The typing methodology that has been presented in Section 3 has been adequately evaluated with respect to its accuracy, yielding very satisfactory results (Maramis et al., 2011). Since HPVTyper employs this methodology, the typing decisions that are made with the help of the application are very accurate. This is true for cases of single HPV infection and – more importantly – for complex cases involving multiple infections. In fact, the aforementioned methodology is the only established way to resolve the latter cases when attempting HPV typing via PCR-RFLP gel electrophoresis.

Automatic typing procedure. HPV typing with HPVTyper can be performed completely automatically. This applies to the entire typing procedure (Sections 3.2 to 3.4) but its importance focuses on the automatic application of the HPV typing algorithm (Section 3.4), since such a feature has been missing from all the previous typing algorithms. In the case of HPVTyper, this feature derives from the computerized nature of the employed methodology along with the use of default values for the set of involved parameters (e.g., parameters θ_1 to θ_{10}) and significantly eases the typing process for the user.

Supervised typing procedure. Although HPVTyper provides the possibility of automatic typing, at the same time, it allows the user to supervise the entire typing procedure. This includes presenting intermediate results to the user (the outcome of initial intensity profile fitting, the outcome of type compatibility check, etc.) and also allowing them to adjust a wide range of parameters that can influence the typing result (e.g., parameters θ_1 to θ_{10}).

This way, the user has full control on the typing procedure and is able to verify themselves step-by-step the correctness of the application's results.

Free availability. Since HPVTyper can be obtained for free, each molecular biology laboratory that possesses the required equipment for performing HPV typing via PCR-RFLP gel electrophoresis can benefit from the presented HPV typing method without any additional cost. In these laboratories the conventional manual method for HPV typing via PCR-RFLP could be replaced with minimum effort by HPVTyper in order to update the employed HPV typing procedure.

Adaptability to emerging discoveries. Scientific research regarding HPV and its connection to cervical cancer is very active. For this reason, it is possible to see in the future new types and genotypes of the virus being discovered, new restriction enzymes and primers being employed for HPV typing, the infection frequency and risk of some HPV types being revised, etc. HPVTyper can easily cope with such scientific discoveries that may emerge by means of its Configuration Editors. The user simply has to edit the information that has become obsolete or add new information through the PCR-RFLP or Type Configuration Editor and the application remains up-to-date. In this sense, HPVTyper is a future-proof HPV typing method.

Expandability to new domains. Although the application has been originally developed as a diagnostic method for the domain of cervical cancer, it can easily find application to every cancer domain in which HPV is involved. Similarly to the previous feature, the information that is related to a new cancer domain can be incorporated in the application with minimum effort by means of the Configuration Editors.

6. Availability

HPVTyper is made freely available to the public through the webpage `http://hippocrates.ee.auth.gr/HPVTyper/` and is subject to the license described therein. The powerful engine module that is installed on the aforementioned server is employed for performing the required computations, while the user interface module is installed to the user's computer by means of the Java Web Start Technology. The use of the aforementioned technology ensures minimum file downloading, automatic updates, and operating system independence at the user's side. A copy of the application's source code can be obtained for free from the above webpage under the same license agreement.

7. Conclusion

In this present chapter, we have introduced HPVTyper, a novel software application for accurate and automatic HPV typing via PCR-RFLP gel electrophoresis. The strength of HPVTyper derives primarily from the underlying HPV typing methodology. This entirely computerized methodology has been justified in parts and as a whole by a series of scientific publications (Maramis & Delopoulos, 2010a;b; Maramis et al., 2010; 2011) and its performance with respect to typing accuracy has been thoroughly evaluated on real HPV typing data (Maramis et al., 2011). By employing this methodology, HPVTyper is able to make accurate typing decisions even for multiply-infected cases in an entirely automatic manner.

The second cornerstone of HPVTyper's strength is its design. The application provides the user with a friendly and straightforward interfacing mechanism to perform all the operations that are involved in the typing process, either as suggested by the application (default parameter values) or according to the user's will (user-adjusted parameter values). Any intermediate results are presented to the user so that they can constantly check the progress of the typing procedure. Owing to the aforementioned design principles, the user has full control over the entire typing process and is able to choose their degree of involvement in the procedure.

The application scope of HPVTyper is impressively wide. By providing tools to easily edit the information that is required for the typing decision, the application can be expanded to include new HPV-related discoveries in the domain of cervical cancer as well as to perform HPV typing in the framework of other cancer domains that are associated with HPV. Finally, it is worth mentioning that the implemented software application could just as well be employed for typing other viruses, should this be required.

A foreseeable future addition to HPVTyper concerns the issue of mutation hotspots in the genome of HPV. When these mutations occur at the restriction sites of the amplified HPV DNA sequence, the resulting FLP of an infecting genotype does not match its expected FLP. Consequently, these mutations constitute a serious threat to the efficacy of the PCR-RFLP typing method in general, no matter whether this is performed manually or by means of HPVTyper.

Fortunately, the powerful decision making mechanism of our application along with the analysis of Section 3.4.2 regarding the phenomenon of partial digestion – already incorporated in the employed methodology – set the ground for future versions of HPVTyper to tackle the aforementioned issue. In fact, these altered mutation-related FLPs are essentially partial digestion FLPs and – just like the latter – they can be foreseen.

With respect to the required modifications on the employed methodology, the compatibility check operation must be updated so as to search for a set of reasonably possible mutation-induced FLPs in addition to the main FLP of an investigated genotype. This flexible compatibility check will allow us to diagnose – with reduced confidence of course (e.g., by adding a penalty to the final score) – infections from HPV types with slightly mutated genomes. However, it should be noted that this approach will induce a significant increase on the computational complexity of the decision making process – an issue that will have to be tackled – and could potentially disturb the accuracy of the employed methodology.

8. References

Barzon, L., Militello, V., Lavezzo, E., Franchin, E., Peta, E., Squarzon, L., Trevisan, M., Pagni, S., Dal Bello, F., Toppo, S. et al. (2011). Human papillomavirus genotyping by 454 next generation sequencing technology, *Journal of Clinical Virology* 52(2): 93–97.

Bosch, F., Lorincz, A., Muñoz, N., Meijer, C. & Shah, K. (2002). The causal relation between human papillomavirus and cervical cancer, *Journal of Clinical Pathology* 55(4): 244–265.

Fontaine, V., Mascaux, C., Weyn, C., Bernis, A., Celio, N., Lefèvre, P., Kaufman, L. & Garbar, C. (2007). Evaluation of combined general primer-mediated PCR sequencing and type-specific PCR strategies for determination of human papillomavirus genotypes in cervical cell specimens, *Journal of Clinical Microbiology* 45(3): 928–934.

Gharizadeh, B., Kalantari, M., Garcia, C., Johansson, B. & Nyrén, P. (2001). Typing of human papillomavirus by pyrosequencing, *Laboratory Investigation* 81(5): 673–679.

Gharizadeh, B., Oggionni, M., Zheng, B., Akom, E., Pourmand, N., Ahmadian, A., Wallin, K. & Nyren, P. (2005). Type-specific multiple sequencing primers: a novel strategy for reliable and rapid genotyping of human papillomaviruses by pyrosequencing technology, *Journal of Molecular Diagnostics* 7(2): 198.

Gheit, T., Landi, S., Gemignani, F., Snijders, P., Vaccarella, S., Franceschi, S., Canzian, F. & Tommasino, M. (2006). Development of a sensitive and specific assay combining multiplex PCR and DNA microarray primer extension to detect high-risk mucosal human papillomavirus types, *Journal of Clinical Microbiology* 44(6): 2025.

Greenblatt, R. (2005). Human papillomaviruses: Diseases, diagnosis, and a possible vaccine, *Clinical Microbiology Newsletter* 27(18): 139–145.

Greer Jr, R., Douglas Jr, J., Breese, P. & Crosby, L. (1990). Evaluation of oral and laryngeal specimens for human papillomavirus (HPV) DNA by dot blot hybridization, *Journal of Oral Pathology & Medicine* 19(1): 35–38.

Husnjak, K., Grce, M., Magdic, L. & Pavelic, K. (2000). Comparison of five different polymerase chain reaction methods for detection of human papillomavirus in cervical cell specimens, *Journal of Virological Methods* 88(2): 125–134.

Hwang, T., Jeong, J., Park, M., Han, H., Choi, H. & Park, T. (2003). Detection and typing of HPV genotypes in various cervical lesions by HPV oligonucleotide microarray, *Gynecologic Oncology* 90(1): 51–56.

Jacobs, M., Snijders, P., Van Den Brule, A., Helmerhorst, T., Meijer, C. & Walboomers, J. (1997). A general primer GP5+/GP6 (+)-mediated PCR-enzyme immunoassay method for rapid detection of 14 high-risk and 6 low-risk human papillomavirus genotypes in cervical scrapings, *Journal of Clinical Microbiology* 35(3): 791.

Jordan, B. (2010). Is there a niche for DNA microarrays in molecular diagnostics?, *Expert Review of Molecular Diagnostics* 10(7): 875–882.

Karlsen, F., Kalantari, M., Jenkins, A., Pettersen, E., Kristensen, G., Holm, R., Johansson, B. & Hagmar, B. (1996). Use of multiple PCR primer sets for optimal detection of human papillomavirus, *Journal of Clinical Microbiology* 34(9): 2095.

Kim, C., Jeong, J., Park, M., Park, T., Park, T., Namkoong, S. & Park, J. (2003). HPV oligonucleotide microarray-based detection of HPV genotypes in cervical neoplastic lesions, *Gynecologic Oncology* 89(2): 210–217.

Klaassen, C., Prinsen, C., De Valk, H., Horrevorts, A., Jeunink, M. & Thunnissen, F. (2004). DNA microarray format for detection and subtyping of human papillomavirus, *Journal of Clinical Microbiology* 42(5): 2152.

Kleter, B., Van Doorn, L., Schrauwen, L., Molijn, A., Sastrowijoto, S., Ter Schegget, J., Lindeman, J., Ter Harmsel, B., Burger, M. & Quint, W. (1999). Development and clinical evaluation of a highly sensitive PCR-reverse hybridization line probe assay for detection and identification of anogenital human papillomavirus, *Journal of Clinical Microbiology* 37(8): 2508.

Landis, S., Murray, T., Bolden, S. & Wingo, P. (1999). Cancer statistics, 1999, *CA: A Cancer Journal for Clinicians* 49(1): 8–31.

Liu, Y. (2008). A technological update of molecular diagnostics for infectious diseases, *Infectious Disorders Drug Targets* 8(3): 183.

Lungu, O., Wright, T. & Silverstein, S. (1992). Typing of human papillomaviruses by polymerase chain reaction amplification with L1 consensus primers and RFLP analysis, *Molecular and Cellular Probes* 6(2): 145–152.

Maramis, C. & Delopoulos, A. (2010a). Efficient quantitative information extraction from PCR-RFLP gel electrophoresis images, *Proc. of 20th International Conference on Pattern Recognition, ICPR 2010*, Istanbul, Turkey, pp. 2564–2567.

Maramis, C. & Delopoulos, A. (2010b). Improved modeling of lane intensity profiles on gel electrophoresis images, *Proc. of XII Mediterranean Conference on Medical and Biological Engineering and Computing, MEDICON 2010*, Porto Karas, Greece, pp. 671–674.

Maramis, C., Delopoulos, A. & Lambropoulos, A. (2010). Analysis of PCR-RFLP gel electrophoresis images for accurate and automated HPV typing, *Proc. of 10th International Conference on Information Technology and Applications in Biomedicine, ITAB 2010*, Corfu, Greece, pp. 1–6.

Maramis, C., Delopoulos, A. & Lambropoulos, A. (2011). A computerized methodology for improved virus typing by PCR-RFLP gel electrophoresis, *IEEE Transactions on Biomedical Engineering* 58(8): 2339–2351.

Mikhailyuk, I. & Razzhivin, A. (2003). Background subtraction in experimental data arrays illustrated by the example of Raman spectra and fluorescent gel electrophoresis patterns, *Instruments and Experimental Techniques* 46(6): 765–769.

Moore, D. (2006). Cervical cancer, *Obstetrics & Gynecology* 107(5): 1152–1161.

Muñoz, N., Bosch, F., de Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K., Snijders, P. & Meijer, C. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer, *New England Journal of Medicine* 348(6): 518–527.

Nobre, R., Almeida, L. & Martins, T. (2008). Complete genotyping of mucosal human papillomavirus using a restriction fragment length polymorphism analysis and an original typing algorithm, *Journal of Clinical Virology* 42(1): 13–21.

Parkin, D. (2006). The global health burden of infection-associated cancers in the year 2002, *International journal of cancer* 118(12): 3030–3044.

Santiago, E., Camacho, L., Junquera, M. & Vázquez, F. (2006). Full HPV typing by a single restriction enzyme, *Journal of Clinical Virology* 37(1): 38–46.

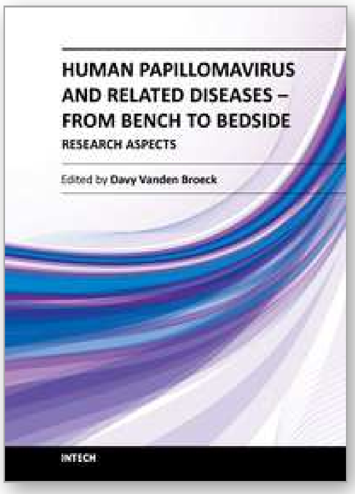
Tagu, D. & Moussard, C. (2006). *Techniques for Molecular Biology*, Science Pub Inc.

van den Brule, A., Pol, R., Fransen-Daalmeijer, N., Schouls, L., Meijer, C. & Snijders, P. (2002). GP5+/6+ PCR followed by reverse line blot analysis enables rapid and high-throughput identification of human papillomavirus genotypes, *Journal of Clinical Microbiology* 40(3): 779.

Vernon, S., Unger, E. & Williams, D. (2000). Comparison of human papillomavirus detection and typing by cycle sequencing, line blotting, and hybrid capture, *Journal of Clinical Microbiology* 38(2): 651.

Vincent, L. & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based onimmersion simulations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(6): 583–598.

Walboomers, J., Jacobs, M., Manos, M., Bosch, F., Kummer, J., Shah, K., Snijders, P., Peto, J., Meijer, C. & Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide, *Journal of Pathology* 189(1): 12–19.



Human Papillomavirus and Related Diseases - From Bench to Bedside - Research aspects
Edited by Dr. Davy Vanden Broeck

ISBN 978-953-307-855-7
Hard cover, 406 pages
Publisher InTech
Published online 25, January, 2012
Published in print edition January, 2012

Cervical cancer is the second most prevalent cancer among women worldwide, and infection with Human Papilloma Virus (HPV) has been identified as the causal agent for this condition. The natural history of cervical cancer is characterized by slow disease progression, rendering the condition, in essence, preventable and even treatable when diagnosed in early stages. Pap smear and the recently introduced prophylactic vaccines are the most prominent prevention options, but despite the availability of these primary and secondary screening tools, the global burden of disease is unfortunately still very high. This book will focus on epidemiological and fundamental research aspects in the area of HPV, and it will update those working in this fast-progressing field with the latest information.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Christos Maramis, Dimitrios Karagiannis and Anastasios Delopoulos (2012). HPVTyper: A Software Application for Automatic HPV Typing via PCR-RFLP Gel Electrophoresis, Human Papillomavirus and Related Diseases - From Bench to Bedside - Research aspects, Dr. Davy Vanden Broeck (Ed.), ISBN: 978-953-307-855-7, InTech, Available from: http://www.intechopen.com/books/human-papillomavirus-and-related-diseases-from-bench-to-bedside-research-aspects/hpvtyper-a-software-application-for-automatic-hpv-typing-via-pcr-rflp-gel-electrophoresis

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821