Accepted Manuscript

Title: Predicting the origin of stains from next generation sequencing mRNA data

Author: Guro Dørum Sabrina Ingold Erin Hanson Jack Ballantyne Lars Snipen Cordula Haas

Please cite this article as: Guro Dorum, Sabrina Ingold, Erin Hanson, Jack Ballantyne, Lars Snipen, Cordula Haas, Predicting the origin of stains from next generation sequencing mRNA data, <![CDATA[*Forensic Science International: Genetics*]]> (2018), https://doi.org/10.1016/j.fsigen.2018.01.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Predicting the origin of stains from next generation sequencing mRNA data

Guro Dørum[a*], Sabrina Ingold [a*], Erin Hanson [b], Jack Ballantyne [b], Lars Snipen [c], Cordula Haas [a]

[*] These authors contributed equally to this work

[a] Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[b] National Center for Forensic Science, University of Central Florida, Orlando, USA

[c] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway

Abstract

We used our previously published NGS mRNA approach for body fluid identification to analyze 183 body fluids/tissues, including mock casework samples. The resulting data set was used to build a probabilistic model that predicts the origin of a stain. Our approach uses partial least squares followed by linear discriminant analysis to classify samples into six commonly occurring forensic body fluids. The model differs from the ones previously suggested in that it incorporates quantitative information (NGS read counts) rather than just presence/absence data. The suggested approach also allows for visualisation of important markers and their correlation with the different body fluids. We compared our model to previously published methods to show that the inclusion of read count information improves the prediction. Finally, we applied the model to mixed body fluid samples to test its ability to identify the individual components in a mixture.

1   Introduction

DNA analysis is widely used to determine who may have left biological evidence at a crime scene. But the composition of a biological stain sample can also be crucial to the investigation and prosecution of a case. A relevant

1