APPLICATIONS IN COMPUTER-ASSISTED BIOLOGY

Harm Nijveen

Thesis committee

Promotor
Prof. Dr A.H.J. Bisseling
Professor of Molecular Biology
Wageningen University

Co-promotor
Dr P.E. van der Vet
Assistant professor, Human Media Interaction Group
University of Twente, Enschede

Other members
Prof. Dr M.A.M. Groenen, Wageningen University
Prof. Dr N.H. Lubsen, Radboud University Nijmegen
Dr J.P.H. Nap, Hanze University of Applied Sciences Groningen
Prof. Dr A.H.C. van Kampen, University of Amsterdam

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences

APPLICATIONS IN COMPUTER-ASSISTED BIOLOGY

Harm Nijveen

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 5 December 2013
at 11 a.m. in the Aula.

Harm Nijveen
Applications in Computer-Assisted Biology
114 pages.

PhD thesis, Wageningen University, Wageningen, NL (2013)
With references, with summaries in Dutch and English

ISBN 978-94-6173-781-6

To the memory of Jack Leunissen

Table of contents

Chapter 1

Introduction

The DNA sequencing revolution

Biology has become a data-rich science, especially due to the massive amount of DNA sequence data that are being collected at still growing rates. Thanks to major advances in the technology (Margulies et al. 2005), the cost of DNA sequencing has dropped steeply in the past decade. The price for sequencing one million nucleotides went from nearly $10,000 in the year 2000 to about 10 cents now (http://www.genome.gov/sequencingcosts). The new DNA sequencing technologies are labelled as next-generation sequencing (NGS), second-generation sequencing, massively parallel sequencing, or high-throughput sequencing to make a distinction with the traditional Sanger sequencing (Sanger et al. 1977). High-throughput sequencing has an enormous effect on all areas of life science research, as the DNA sequence is such a fundamental property in biology. DNA encodes the amino acid sequence of the cell's proteins, but also the developmental program of when and at what rate each protein should be produced, in a way that we are only just beginning to understand (Encode Project Consortium et al. 2007).

After completion of the human genome project (Collins et al. 2003), similar projects followed in its wake, and to date the genomes of nearly 7,000 species have been sequenced, including 300 eukaryotes(Pagani et al. 2012). Completing the sequence of a species' genome, although an important milestone, is usually only the starting point for many applications of DNA sequencing technology in biological research. Sequencing many individuals of the same species provides information about the nucleotide variation within that species (Gan et al. 2011; 1000 Genomes Project Consortium et al. 2012). Gene expression under various conditions can be studied by analysing the messenger RNAs that are transcribed (transcriptomics) (Wang et al. 2009). DNA-protein interactions can be mapped with chromatin immunoprecipitation-sequencing (ChIP-Seq) (Shendure and Ji 2008). And still new approaches involving DNA sequencing are being developed, addressing very diverse biological questions in ways that were not feasible only a few years ago. A similar, albeit less revolutionary development is taking place with the large-scale identification of proteins and metabolites using mass spectrometry (Aebersold and Mann 2003; De Vos et al. 2007).

The amount of data that is generated by a typical high-throughput DNA sequencing experiment is several orders of magnitude higher than with Sanger sequencing. Where previously a researcher examined a small well-defined set of genes, now thousands of genes in a genome or even several genomes can be studied at the same time. This dramatically speeds up biological research and enables new approaches to study biological systems. The genome scale of the data that are produced in an experiment also makes it potentially useful for other researchers that are interested in their own 'pet' genes or biological pathways. To accommodate the reuse of NGS data, the Sequence Read Archive (SRA)

(Kodama et al. 2012) has been established as part of the International Nucleotide Sequence Database Collaboration (INSDC). This INSDC is a partnership of the three main nucleotide sequence databases ENA, GenBank, and DDBJ (Nakamura et al. 2013). The SRA is a repository for storing raw high-throughput sequencing data annotated with details of the experiment they were obtained in. It currently already contains over 1,500 terabases (http://www.ncbi.nlm.nih.gov/Traces/sra/), which compares in size to about half a million times a single human genome.

From data to information

As a result of the richly available biological (sequence) data the computer has become an indispensable tool for biological research, and biologists are spending more and more time behind the computer keyboard. However, extracting useful information out of the large volumes of complex data to gain knowledge has now become the challenging part of the experiment (Marx 2013), a situation adequately described by Elaine Mardis as "the $1,000 genome, the $100,000 analysis?" (Mardis 2010; Sboner et al. 2011). Manual analysis of the data is no longer feasible, and standard office software is often not suitable, even for inspecting the data. This has strongly boosted the demand for software tools aimed at biological research and many of these tools have been developed in the recent years for a wide array of different biological analyses.

Most of the current biological software tools were made by researchers who needed them for their own research and then decided to make the software available to the community, often through an "application note" publication in a scientific journal. As a result, these applications vary hugely in terms of robustness, implementation and usability. Many of the tools have to be operated via the (Linux) command-line, and therefore require advanced computer skills to be used. It is also not uncommon that multiple tools are available for a specific task, as is the case for genome assembly software (Zhang et al. 2011). This apparent redundancy is probably an indication that there is not one single tool yet that adequately addresses all of the different aspects of a complex task.

Biological databases

Next to software tools the number of biological databases is also increasing rapidly. The Nucleic Acids Research (NAR) online Molecular Biology Database Collection already contains over 1,500 different publicly available databases (Fernandez-Suarez and Galperin 2013). These databases contain information about biological properties as diverse as transcription factor binding sites, protein domains and protein-protein interactions. The databases are often made accessible with a set of supporting web-based tools that allow querying or summarizing the data. In contrast to the above discussed software tools, these database linked web interfaces are usually much more accessible for biologists that do not have expert computer skills. A popular web site that hosts many biological databases is provided by the National Center for Biotechnology Information (NCBI) (Ncbi

Resource Coordinators 2013). It allows easy searching and retrieval of all kinds of information from nucleic acid sequences in GenBank (Benson et al. 2013) to single nucleotide variations in dbSNP (Sherry et al. 2001) and literature in PubMed. At the other side of the Atlantic Ocean the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger institute at the Wellcome Trust Genome Campus provide similar resources (Goujon et al. 2010; Flicek et al. 2013).

Bioinformatics

The field of bioinformatics can be characterized as "… the application of computational techniques to understand and organise the information associated with biological macromolecules." (Luscombe et al. 2001). As such of all biological disciplines, bioinformatics appears best suited to deal with large volumes of biological data. Exactly the topics discussed above, the development and use of software and biological database are the domain of bioinformaticians, working to extract information from data in order to gain knowledge. This could be considered the golden age of bioinformatics.

Thesis outline

This thesis discusses the development and application of software solutions in biological research. Chapter 2 describes Primer3Plus, a web application for designing oligonucleotide primers that can for instance be used in PCR experiments. Oligonucleotide primers, or primers, are short DNA molecules that bind to a DNA or RNA template and as such serve as the starting point for DNA duplication by a DNA polymerase. The sequence of the primer determines which part of the template will be amplified, provided that the primer binds specifically to its complementary sequence. The optimal temperature for specific binding of a primer to the template depends on the nucleotide sequence of the primer and experimental conditions like the magnesium concentration. A simple back-of-the-envelope formula for calculating the optimal temperature for a short DNA sequence is the Wallace-Ikatura rule (Suggs et al. 1981) which adds two degrees for an adenosine or a thymine and four degrees for each cytosine or guanine:

$$\mathrm{Tm}\ (^{\circ}\mathrm{C}) = 2\ (\mathrm{A}+\mathrm{T}) + 4\ (\mathrm{G}+\mathrm{C})$$

Work by Santa Lucia based on the actual melting temperature of many different oligonucleotides provides a much more sophisticated method for calculating the melting temperature (SantaLucia 1998), but the complexity of this method introduces the need for a computer. One of the most used software tools to design primers is Primer3 (Rozen and Skaletsky 1999). We developed Primer3Plus as an enhanced task oriented web interface to Primer3. Since its publication in 2007, Primer3Plus is used from more than 175,000 different unique internet addresses and cited by hundreds of peer reviewed articles, showing that it fulfils a clear

need in the biological research community.
Chapter 3 discusses QualitySNPng, a software tool for finding single nucleotide polymorphisms (SNPs) in high-throughput sequencing data. Biological variation in the form of SNPs is important for linking phenotypic traits to genes, as is done in genome wide association studies (GWAS) and quantitative trait locus (QTL) analysis (Davey et al. 2011; Nielsen et al. 2011). One of the difficulties with finding SNPs in DNA and RNA sequencing data is distinguishing the true biological variation from sequencing errors. That is a task where SNP detection software can help. QualitySNPng is a user-friendly software tool that uses several criteria for detecting SNPs in high throughput sequencing data and allows the biologist to modify algorithm parameters based on previous knowledge of the biological system under study. It was inspired on the QualitySNP pipeline for SNP detection (Tang et al. 2006). Based on the detected SNPs, QualitySNPng predicts haplotypes from alleles that are linked by sequence reads. These haplotypes can serve to identify genotypes from the sequenced individuals. This genotyping-by-sequencing (GBS) (Davey et al. 2011) combines marker (SNP) detection with marker scoring and can thereby hugely speed up for instance precision breeding approaches.

Chapter 4 is about the Human Short Peptide Variation database (HSPVdb) that was developed together with mass spectrometry and immunology experts from the Leiden University Medical Center. The database was created as a resource for human peptides, including the variations that can arise from SNPs. In the human population there exist many nucleotide variations that are not apparently harmful to their bearers, even if these SNPs lead to single amino acid variations (SAPs) and thus change the protein sequence. This natural variation in human proteins is not reflected by the protein sequences in the main public protein databases like UniProt (UniProt Consortium 2013), since that only contains one variant of the protein. These variations in protein sequence might not have a strong effect on protein function, they are important in the practice of tissue transplantation, since even a single amino acid difference between a protein from the donor and the patient can elicit a strong immune response. The HSPVdb was created for a project to find peptides that can trigger a so-called graft-versus-leukaemia effect (Marijt et al. 2003), which is a beneficial form of the well-known graft-versus-host effect where the immune response is specifically directed against tumour cells. The HSPVdb combines human protein sequence information with known nucleotide variations to produce an extensive list of possible human peptides. Each peptide is annotated with the SNPs that are contained in its coding sequence, and their corresponding frequencies in the human population. The database can be queried from a web interface that allows searching for peptide fragments. This web tool combines several steps that the biologist previously did by hand: search with the peptide sequence in the human protein database using BLAST, find any known SNPs in the region of the protein that the peptide was derived from, analyse

the effect of these SNPs on the amino acid sequence. This data integration not only saves the biologist from performing a time consuming repetitive task, it also avoids the introduction of human errors during the manual data collection steps.

The work in chapter 5a was performed in collaboration with a microbiologist interested in finding mobile promoters in bacteria. Promoter duplication was recently shown to be an important step for the evolution of an *Escherichia coli* population to grown on citrate in the long-term evolution experiment (Blount et al. 2012). We created a conservative inventory of mobile promoters within the genomes of all publicly available bacterial genomes by comparing all promoters within a genome with each other. We used BLAST (Altschul et al. 1997) for the comparison and netclust (Kuzniar et al. 2010) for the subsequent clustering of the pairwise hits. Various other publicly available databases and software tools were used for additional studies, like functional enrichment analysis. Chapter 5b is a commentary on the work in chapter 5a, that elaborates on mobile promoters that are shared between distantly related species.

The developments in high-throughput technologies have such a revolutionizing effect on life science research that the field of biology is fundamentally changing, leading to the emergence of New Biology. This New Biology strongly integrates biology with other sciences to address challenging medical and societal problems. Biology is also turning into a much more quantitative science and biological researchers will have to adapt to their changing environment. Chapter 6 discusses the road ahead for the integration of computational methods into biological research. I propose three approaches that are complementary. In the first approach the biologist is empowered through user-friendly software, allowing him to do standardized computational analyses without expert computer skills. In the second approach the biologist collaborates with a bioinformatician to combine in-depth biological knowledge with expert computational skills. For the third approach the biologist himself learns to do the bioinformatics analysis, thereby combining the biologist and the bioinformatician in one person. In addition to these three approaches, it seems evident that every biologist will have to learn at least the basics of computational data analysis. Biological research will only become more data-rich and a large part of most research projects will be spent doing, essentially, bioinformatics.

REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* **422**(6928): 198-207.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res* **41**(Database issue): D36-42.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**(7417): 513-518.

Collins FS, Morgan M, Patrinos A. 2003. The Human Genome Project: lessons from large-scale biology. *Science* **300**(5617): 286-290.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**(7): 499-510.

De Vos RC, Moco S, Lommen A, Keurentjes JJ, Bino RJ, Hall RD. 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Prot* **2**(4): 778-791.

Encode Project Consortium Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Fernandez-Suarez XM, Galperin MY. 2013. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res* **41**(Database issue): D1-7.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41**(Database issue): D48-55.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**(7365): 419-423.

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* **38**(Web Server issue): W695-699.

Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C. 2012. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **40**(Database issue): D54-56.

Kuzniar A, Dhir S, Nijveen H, Pongor S, Leunissen JA. 2010. Multi-netclust: an efficient tool for finding connected clusters in multi-parametric networks. *Bioinformatics* **26**(19): 2482-2483.

Luscombe NM, Greenbaum D, Gerstein M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* **40**(4): 346-358.

Mardis ER. 2010. The $1,000 genome, the $100,000 analysis? *Genome Med* **2**(11): 84.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marijt WA, Heemskerk MH, Kloosterboer FM, Goulmy E, Kester MG, van der Hoorn MA, van Luxemburg-Heys SA, Hoogeboom M, Mutis T, Drijfhout JW et al. 2003. Hematopoiesis-restricted minor histocompatibility antigens HA-1- or HA-2-specific T cells can induce complete remissions of relapsed leukemia. *Proc Natl Acad Sci U S A* **100**(5): 2742-2747.

Marx V. 2013. Biology: The big challenges of big data. *Nature* **498**(7453): 255-260.

Nakamura Y, Cochrane G, Karsch-Mizrachi I, International Nucleotide Sequence Database C. 2013. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* **41**(Database issue): D21-24.

Ncbi Resource Coordinators. 2013. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **41**(Database issue): D8-D20.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**(6): 443-451.

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**(Database issue): D571-579.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**:365-386

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12): 5463–5467.

SantaLucia J, Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**(4): 1460-1465.

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biology* **12**(8): 125.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**(10): 1135-1145.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311.

Suggs SV, Hirose T, Miyake EH, Kawashima MJ, Johnson KI, Wallace RB. 1981. Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences. *ICN-UCLA Symp. Dev. Biol.* **23**:683-693.

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA. 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* **7**: 438.

UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* **41**(Database issue): D43-47.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* **6**(3): e17915.

Chapter 2

Primer3Plus, an enhanced web interface to Primer3

Nucleic Acids Research, 2007, **35**(Web Server issue): W71-74.

Andreas Untergasser & Harm Nijveen, Xiangyu Rao, Ton Bisseling, René Geurts, Jack A. M. Leunissen

Andreas Untergasser and Harm Nijveen share first authorship

ABSTRACT

Here we present Primer3Plus, a new web interface to the popular Primer3 primer design program as an enhanced alternative for the CGI-scripts that come with Primer3. Primer3 consists of a command line program and a web interface. The web interface is one large form showing all of the possible options. This makes the interface powerful, but at the same time confusing for occasional users. Primer3Plus provides an intuitive user interface using present-day web technologies and has been developed in close collaboration with molecular biologists and technicians regularly designing primers. It focuses on the task at hand, and hides detailed settings from the user until these are needed. We also added functionality to automate specific tasks like designing primers for cloning or step-wise sequencing. Settings and designed primer sequences can be stored locally for later use. Primer3Plus supports a range of common sequence formats, such as FASTA. Finally, primers selected by Primer3Plus can be sent to an order form, allowing tight integration into laboratory ordering systems. Moreover, the open architecture of Primer3Plus allows easy expansion or integration of external software packages. The Primer3Plus Perl source code is available under GPL license from SourceForge.

Primer3Plus is available at http://www.bioinformatics.nl/primer3plus

INTRODUCTION

Oligonucleotide primers are widely used in various molecular biology techniques like DNA sequencing and the polymerase chain reaction (PCR). Since a primer serves as the starting point for DNA replication, specific binding of the oligonucleotide to the target sequence on the template strand is essential for a successful experiment. The binding specificity of a primer is determined by several of its properties, like the melting temperature (Tm), GC-content and self complementarity. Designing primers is usually done with the help of computer programs, among which Primer3 is most widely used judging from the hundreds of citations of the primary publication (Rozen and Skaletsky 2000). Primer3 is popular since the program can be used online and is redistributed free of charge. Recently, a SourceForge project was started for Primer3 in which several improvements are being discussed and implemented. Subsequently Primer3 was updated to include an additional method for Tm calculation [Santa Lucia mode, (Koressaar and Remm 2007)], since it was argued that primer design based on the Breslauer model shows variation in the predicted Tm (Breslauer et al. 1986; Gordon and Sensen 2004).

Primer3 consists of a command line C program, and an HTML web interface written in Perl. The web interface is one large form showing all possible options. This makes it powerful, but at the same time confusing for the occasional user; without in-depth knowledge it is hard to tell which of the settings are important for a specific design task. Some tasks, like designing PCR primers for position specific cloning, are not easy to perform with the current web interface, since this

would require multiple runs of the program under different settings. Because of these issues with the current web interface and facilitated by its liberal license, several academic and commercial parties have created their own command line and/or web interface to Primer3, tailored to their specific needs and audience; e.g. EMBOSS' EPRIMER3 and the PCR Suite (Rice et al. 2000; van Baren and Heutink 2004). We took this effort one step further and developed a general purpose, easy to use and powerful new web interface for Primer3, called Primer3Plus that is available at http://www.bioinformatics.nl/primer3plus

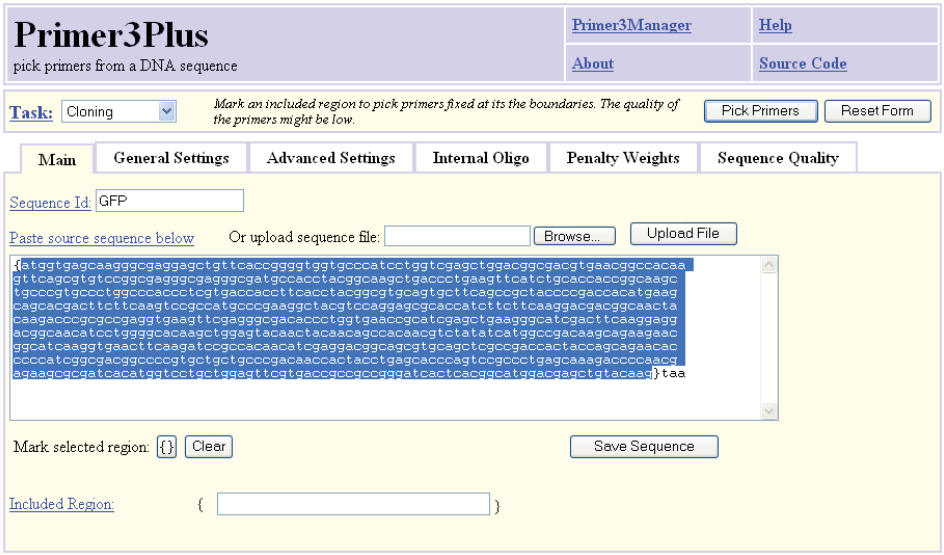


Figure 1. Screenshot of the Primer3Plus start page. In contrast to the original primer3 web interface, detailed settings that are not frequently used are hidden behind tabbed panels. A pull-down task menu is added (top left) that provides 5 scenarios for which Primer3Plus can be used: gene detection, primers for cloning purposes, sequencing, list of a possible primers and check option of already available primers. Here an example sequence (GFP) was loaded, the Cloning task selected and the open reading frame marked as included region.

Web interface

We aimed to design the Primer3Plus web interface in such a way that it is comprehensible for occasional users, and yet powerful for the experienced users that need to do more complex or laborious tasks. Where the original Primer3 web interface presents all options and settings in one large web page, Primer3Plus starts with a more simple screen containing only the input boxes for the sequence and the option to select the target region for amplification (Figure 1). The target sequence can be pasted or uploaded in any common format, such as FASTA and EMBL. Furthermore, the "Pick Primers" button is moved to the top of the page to make sure it is always available. Also a new task selection box is included (top left corner), that provides the possibility to select between 5 different scenarios

for which Primer3Plus can be used (see subsequently).
For most purposes the default parameter settings are adequate. Adjusting these parameters is possible by using the tabbed panels. The tabs are ordered in the most likely frequency of use, with the left most panel being the start page, followed by the general settings, advanced settings, internal oligo options, penalty weights and sequence quality. The parameter names were kept the same as in the original Primer3 interface to minimise time needed to get familiar with using Primer3Plus. The labels of the parameters are web links to a context-sensitive help text, just like in the original Primer3 web interface. Primer3Plus adds to this a floating tool tip with a brief description for some of the most prominent parameters. Next to the default settings, the user can choose specific settings for special tasks like primer design for qPCR applications. It is also possible to store custom settings locally to save the user from having to go through the configuration each time primers are designed for specific non-standard conditions, like AT-rich organisms or high-salt environments.
The Primer3Plus results page shows the suggested primers with their characteristics, ordered from best to worst. The best scoring primer pair is marked on the template sequence. The desired primers can be selected and submitted to Primer3Manager where the user can manage a primer collection. Designed primers are stored on the web server for a limited period, allowing users to design multiple primer pairs and combine them in a single order. Users, or actually web clients, are identified using cookies that expire after a set period (currently 1 week). Primers can also be saved as a FASTA file and previously saved primers can be uploaded allowing the addition of new primers to existing primer collections. Within the Primer3Manager selected primers can be re-analysed, submitted to the NCBI BLAST service (Altschul et al. 1997) or combined on an order form. In case Primer3Plus is run on a local server, this order form can be customized allowing tight integration into a laboratory primer order management system.

Design tasks

We discriminate five distinct primer design user scenarios, or tasks, that have been integrated in the new interface. These tasks can be selected using a drop down list at the top of the Primer3Plus page. Currently the tasks include Detection, Cloning, Sequencing, Primer Check and Primer List. More tasks might be added in future versions of the Primer3Plus web interface.

Detection

The Detection task can be used for designing standard PCR primers or hybridisation oligos to detect a given sequence. The user provides the template sequence and optionally indicates target regions within the sequence or changes the parameters for primer picking. By pressing the Pick Primers button a list of suggested primer pairs is presented, ordered with the best pair at the top. The Detection task is most reminiscent to the original Primer3 web interface.

Cloning
For cloning a PCR product in a specific reading frame in an expression vector, it is important to control precisely the start and end of the PCR product. With the Cloning task either the 5' or the 3' ends of the primers can be fixed to the boundary of the included region, marked with curly braces. The output lists the predicted best fitting primer pairs that start or end at the given position. Subsequently, the user can modify the primer sequence for example, e.g. adding a restriction enzyme recognition site that facilitates cloning of the PCR product. Fixing primer ends significantly limits the number of possible candidates that fulfil the stringent default settings. If this is the case, alternative primers are given that are accompanied by a warning concerning the relaxed stringencies that have been applied (Figure 2).

Sequencing
The Sequencing task is developed to design a series of primers on both the forward and reverse strands that can be used for custom primer-based (re-) sequencing of clones. Under default settings, the primers are spaced by a 500 nt interval on both strands. The spacing of primers is configurable within the Advanced settings tab, as is the overlap between the forward and reverse primers. A configurable number of nucleotides immediately following the primer that usually cannot be read reliably, are taken into account when selecting the forward and reverse primers.

Primer Check
The Primer Check task can be used to obtain information on a specified primer, like its melting temperature or self complementarity. This can, for instance, be useful when certain primers are to be reused under different experimental conditions or if the original primer characteristics were lost. The Primer Check option is also available on the Primer3Manager page, enabling a re-check of primers that have been modified by the user, e.g. for cloning purposes. When this task is selected, no template sequence is required and only the primer sequence has to be provided.

Primer List
With the Primer List task all possible primers that can be designed on the target sequence and meet the current settings will be returned with their corresponding characteristics. This task basically allows manual selection of primers, which could be of interest in case of specific tasks that are not (yet) implemented in Primer3Plus (like picking overlapping primers). The run time of the Primer List task can be relatively long when compared to the other Task options, and can take up to a minute, especially when lengthy target sequences are submitted.

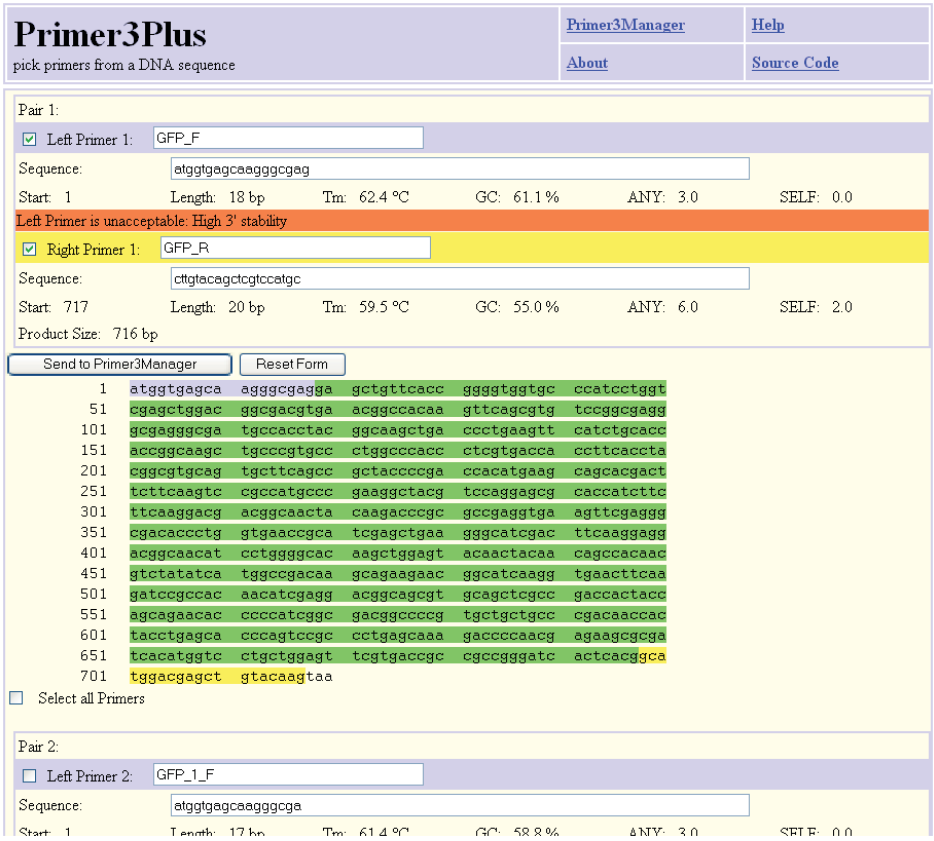


Figure 2. Result page of a Primer3Plus Cloning run showing the left and right primers in blue and yellow. The included region is coloured green. Although the left primer is of low quality due to high 3' stability, it is the best primer starting at this position with the given settings. Alternative primer pairs of lower quality are shown below the sequence (truncated here).

Technology

The Primer3Plus web pages consist of HTML with JavaScript for interactivity. The JavaScript is not essential for proper function of Primer3Plus and can be disabled. Custom settings are stored locally using the Boulder IO format (a simple data format: tag=value). The Primer3Plus web interface has been tested to run properly on the most recent versions of Firefox, Internet Explorer, Safari and Konqueror. The server side software has been written in Perl, using the CGI standard for communicating with the Apache web server. It runs on both SuSE Enterprise Linux 9 and Windows XP, and should run unaltered on any modern Unix/Linux platform.

Primer3Plus is designed with interoperability in mind. It is relatively easy to integrate with other (web) applications. Settings, like the template sequence, can

be passed in via HTML POST or GET operations, so applications can be linked directly to an already filled Primer3Plus form.

FUTURE WORK

An annotated view of the template DNA sequence showing, for instance, the open reading frames or intron/exon boundaries would facilitate finding the regions of interest for amplification or detection. For this, Primer3Plus will need to be extended with functionality to extract and display features from uploaded sequence files.

Primer3Plus was recently included in the Primer3 SourceForge project (http://sourceforge.net/projects/primer3/). This allows for tight integration of the web interface with the core program and creates a platform for both developers and end users to discuss further modifications and additions to Primer3Plus.

ACKNOWLEDGEMENT

The authors would like to thank Steve Rozen for helpful suggestions and ongoing support of Primer3, and Catarina Cardoso and Gerben Bijl for extensive beta testing.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Breslauer KJ, Frank R, Blocker H, Marky LA. 1986. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* **83**(11): 3746-3750.

Gordon PM, Sensen CW. 2004. Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays. *Nucleic Acids Res* **32**(17): e133.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**(10): 1289-1291.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**(6): 276-277.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**:365-386.

van Baren MJ, Heutink P. 2004. The PCR suite. *Bioinformatics* **20**(4): 591-593.

Chapter 3

QualitySNPng: a user-friendly SNP detection and visualisation tool

Nucleic Acids Research, 2013, **41**(Web Server issue): W587-590.

Harm Nijveen, Martijn van Kaauwen, Danny G. Esselink, Brechtje Hoegen, Ben Vosman

ABSTRACT

QualitySNPng is a new software tool for the detection and interactive visualisation of single nucleotide polymorphisms (SNPs). It uses a haplotype-based strategy to identify reliable SNPs, is optimized for the analysis of current RNA-seq data, but can also be used on genomic DNA sequences derived from next generation sequencing experiments. QualitySNPng does not require a sequenced reference genome and delivers reliable SNPs for di- as well as polyploid species. The tool features a user-friendly interface, multiple filtering options to handle typical sequencing errors, support for SAM and ACE files and interactive visualisation. QualitySNPng produces high quality SNP information that can be used directly in genotyping by sequencing approaches for application in QTL and genome wide association mapping as well as to populate SNP arrays. The software can be used as a stand-alone application with a graphical user interface or as part of a pipeline system like Galaxy. Versions for Windows, Mac OS X and Linux as well as the source code are available from**:** http://www.bioinformatics.nl/QualitySNPng

INTRODUCTION

Recent developments in sequencing technology have revolutionized genetic research as vast amounts of sequencing data are now becoming available. From this data SNP information can be extracted that is useful for genetic analysis, including QTL mapping and genome wide association studies (Davey et al. 2011; Nielsen et al. 2011). Although several tools for SNP detection are already available (Li et al. 2009; Koboldt et al. 2009; DePristo et al. 2011) they usually require Linux command line skills to run and use of a separate program to visualise the results. More user-friendly software would greatly benefit the community.

Since its publication the QualitySNP pipeline for SNP detection in diploid and polyploidy species(Tang et al. 2006) has been successfully used in dozens of projects in plant and animal genetics, for instance for the identification of SNP markers in crop plants (Anithakumari et al. 2010), zebra finch (Stapley et al. 2008), waterfleas (Orsini et al. 2011), snakes (Cardoso et al. 2010) and scallops (Hou et al. 2011). Because QualitySNP can use *de novo* assembled sequence alignments as input it can also be used for species without a reference genome. The original QualitySNP was developed and optimized for Sanger sequenced expressed sequence tag (EST) data; however, the nature of DNA and RNA sequencing has changed drastically over the past six years, making an update necessary. Here we present QualitySNPng that was specifically tuned to identify SNPs in data from the current next generation sequencing platforms. It features a graphical user interface (GUI), supports the popular SAM format (Li et al. 2009), general performance improvements to allow analysis of large data sets and additional filtering parameters that address specific characteristics of NGS data from different platforms. The identified SNPs can be viewed in the context of predicted haplotypes and per input sample, making it ideally suited

for a genotyping by sequencing approach (Davey et al. 2011). Additionally QualitySNPng can be used as a component in an analysis pipeline like the Galaxy platform (Goecks et al. 2010).

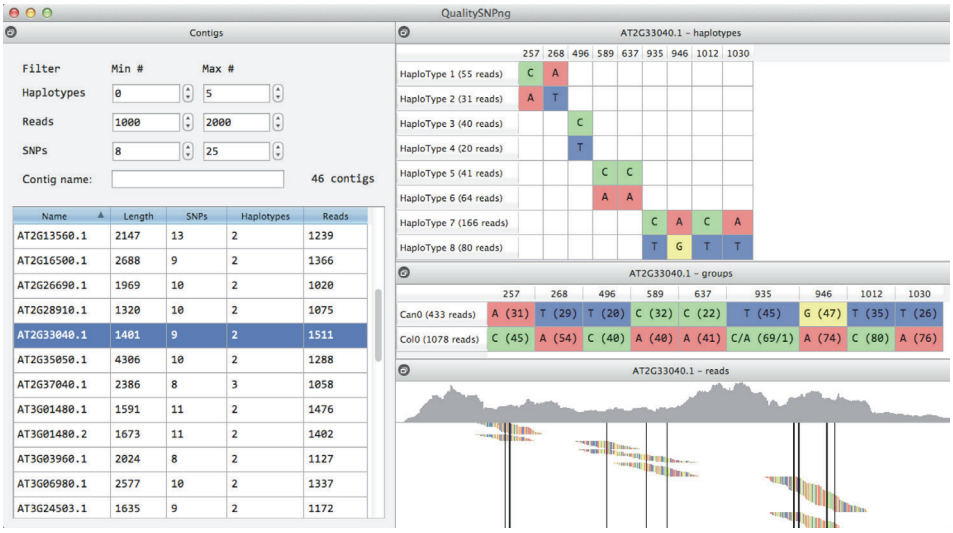


Figure 1. Screenshot of QualitySNPng output. Result of the SNP detection using Arabidopsis thaliana RNA-seq dataset from two accessions that were mapped to Arabidopsis transcripts (Gan et al. 2011). In the left panel the list of transcripts is shown, limited here using the filter options to only the ones with between 8 and 25 SNPs and between 1000 and 2000 reads. The details for the selected transcript are shown on the right: the top window shows the predicted haplotypes, the middle window shows the alleles per accession (Col-0 and Can-0), and the bottom window shows the reads aligned to the transcript sorted per haplotype (reads without SNP are not shown).

FEATURES

SNP calling

QualitySNPng takes as input a sequence alignment file in SAM (Li et al. 2009) or ACE (Gordon et al. 1998) format with single-end or paired-end reads as produced by read mappers like Bowtie (Langmead et al. 2009) and BWA (Li and Durbin 2009) or *de novo* assemblers like CABOG (Miller et al. 2008) and PCAP (Huang et al. 2003). The QualitySNPng software employs three filtering steps to eliminate unreliable variations similar to the original QualitySNP (Tang et al. 2006). The first filter labels all nucleotide differences that occur in a minimum number of reads as *potential* SNPs. This minimum number can be adjusted by the user as an absolute number or a fraction of the total number of reads. The second filter takes into account the quality of the sequence containing the variant nucleotide and leaves only the *high confidence* SNPs. The base quality, characterized by the Phred score (Ewing et al. 1998), is used for this when it is present in the input sequence alignment. If no Phred score is present, all nucleotides in the input reads are assumed to be of high quality. Additionally, the score can be modified based

on specific sequence patterns. For instance, variations found in homopolymeric tracts can be set to low quality. This option is particularly useful when Roche/454 sequences are processed as these are known to be prone to homopolymer-associated errors (Margulies et al. 2005). Also a number of nucleotides at the 5' or 3' ends can be labelled as low quality, for instance to avoid false SNPs caused by incomplete adaptor trimming. The third filter involves predicting haplotypes based on the high confidence SNPs. Only if variation is supported by one or more haplotypes it is considered as a *reliable SNP*. Compared to the original QualitySNP software, the second and third filters were reversed to make sure that the detected haplotypes are based on high confidence SNPs only. The run time largely depends on the size and nature of the input sequencing data, ranging from less than a minute for a set of ~25,000 contigs (~100 reads/contig), to 10 minutes for one large single contig of 7,000 bp with 800,000 reads. Larger and more variable sequence alignments can take longer, also depending on the stringency of the settings: lowering the threshold for potential SNPs will result in more work for the second and third filters that are computationally the most expensive. For large input files that are expected to take several hours to process one can use the command line 'server mode' option of the tool to perform the SNP calling on a compute server and subsequently analyse the results using the GUI.

Viewing results

The results of the SNP calling can be viewed directly using the GUI, and are also saved in structured text files for later reference or further processing. The different contigs from the input sequence alignments are listed in a table showing the number of SNPs, the reads and the haplotypes. The haplotype count in the table is corrected for fragmented haplotypes by taking the maximal number of haplotypes that is found per SNP position. Fragmentation of haplotypes may occur and is caused by SNPs that are too far apart to be linked to one allele by a single sequence read or a read pair, see Figure 1 for an example. The contig list can be filtered based on the numbers of reads, SNPs and haplotypes and (partial) contig name.

A selected contig will show a window with the aligned reads and the SNPs indicated, a table with the haplotypes and their alleles per SNP position, and a table showing the alleles for the different samples in the input data (Figure 1). For this last table to appear, the input sequence alignment file should be annotated with a "read group" (see SAM format definition) per read, or alternatively, have group labels included in the read names. The overview per sample can for instance be used to compare alleles between different accessions, strains or ecotypes and for genotyping by sequencing.

Manual inspection of the read alignment together with the haplotype overview gives insight in the quality of the alignment, local read coverage and positions of the SNPs. Based on this visual inspection one can decide to alter the stringency

of the filter settings and rerun the SNP calling. The reads can be sorted on start position or per haplotype, and viewed at different zoom levels.

For the creation of a SNP array, marker SNPs can be selected and exported with flanking sequence of a specified length as a structured text file that can be imported into a standard spread sheet program or an assay design program.
To avoid problems in SNP scoring we suggest to select markers from contigs that have no more than the maximum expected number of haplotypes, i.e. two for diploid species, as contigs with more haplotypes may contain paralogous sequences. To further increase the chance of obtaining markers that will perform well on arrays one could use the BLAST program (Altschul et al. 1997) to eliminate marker sequences that show high similarity to other genes, as was shown previously (Anithakumari et al. 2010).

IMPLEMENTATION

QualitySNPng was written in C++ using the Qt toolkit. The same executable file can be used interactively with the graphical user interface, or as a command line tool for inclusion in analysis pipelines to be run on a compute server. The software can be compiled and runs on the Windows, Mac OS X and Linux operating systems. The output data is saved as CSV text files and can be reloaded for later analysis using QualitySNPng, or processed by custom scripts for further analysis.

DISCUSSION AND FUTURE DIRECTIONS

We believe there is a strong need for user-friendly software tools that allow biologists to directly analyse and visualise their data. QualitySNPng is versatile tool that combines SNP detection and genotyping with interactive visualisation of the results. The GUI with its pre-set filter options is easy to use, and also highly configurable for specific needs. QualitySNPng is routinely used in-house for marker SNP identification in several projects (Golas et al. 2013; Shahin et al. 2012a; 2012b). In one project QualitySNPng was used to analyse RNA-seq data with up to 8 million reads per transcript to genotype a mixture of a few hundred accessions (unpublished) by making use of the ‘server mode’ option to run on a compute server. We expect that developments like in cloud computing will make this possible without leaving the graphical user interface. The source code of QualitySNPng is freely available and we encourage further development and implementation of the software in custom SNP analysis pipelines or adaptation for specific applications.

ACKNOWLEDGEMENTS

We thank Thomas van Gurp for valuable feedback during the development process. This work is dedicated to the memory of Professor Jack A.M. Leunissen, who sadly passed away in May 2012 and was one of the initiators of this project.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Anithakumari AM, Tang J, van Eck HJ, Visser RGF, Leunissen JAM, Vosman B, van der Linden CG. 2010. A pipeline for high throughput detection and mapping of SNPs from EST databases. *Mol Breed* **26**: 65–75.

Cardoso KC, Da Silva MJ, Costa GGL, Torres TT, Del Bem LEV, Vidal RO, Menossi M, Hyslop S. 2010. A transcriptomic analysis of gene expression in the venom gland of the snake Bothrops alternatus (urutu). *BMC Genomics* **11**: 605.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Ewing BB, Hillier LL, Wendl MCM, Green PP. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genes Dev* **8**: 175–185.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**: 419–423.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**: R86.

Golas TM, van de Geest H, Gros J, Sikkema A, D'Agostino N, Nap JP, Mariani C, Allefs JJHM, Rieu I. 2013. Comparative next-generation mapping of the Phytophthora infestans resistance gene Rpi-dlc2 in a European accession of Solanum dulcamara. *Theor Appl Genet* **126**: 59–68.

Gordon DD, Abajian CC, Green PP. 1998. Consed: a graphical tool for sequence finishing. *Genes Dev* **8**: 195–202.

Hou R, Bao Z, Wang S, Su H, Li Y, Du H, Hu J, Wang S, Hu X. 2011. Transcriptome sequencing and de novo analysis for Yesso scallop (Patinopecten yessoensis) using 454 GS FLX. *PLoS ONE* **6**: e21560.

Huang X, Wang J, Aluru S, Yang S-P, Hillier L. 2003. PCAP: a whole-genome assembly program. *Genome Res* **13**: 2164–2170.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283–2285.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**: 2818–2824.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Orsini L, Jansen M, Souche EL, Geldof S, De Meester L. 2011. Single nucleotide polymorphism discovery from expressed sequence tags in the waterflea Daphnia magna. *BMC Genomics* **12**: 309.

Shahin A, van Gurp T, Peters SA, Visser RG, van Tuyl JM, Arens P. 2012a. SNP markers retrieval for a non-model species: a practical approach. *BMC Res Notes* **5**: 79–79.

Shahin A, van Kaauwen M, Esselink D, Bargsten JW, van Tuyl JM, Visser RG, Arens P. 2012b. Generation and analysis of expressed sequence tags in the extreme large genomes Lilium and Tulipa. *BMC Genomics* **13**: 640–640.

Stapley J, Birkhead TR, Burke T, Slate J. 2008. A linkage map of the zebra finch Taeniopygia guttata provides new insights into avian genome evolution. *Genetics* **179**: 651–667.

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JAM. 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* **7**: 438.

Chapter 4

HSPVdb – the Human Short Peptide Variation database for improved mass spectrometry-based detection of polymorphic HLA-ligands

Immunogenetics, 2011, **63**(3): 143-153.

Harm Nijveen & Michel G.D. Kester, Chopie Hassan, Aurélie Viars, Arnoud H. de Ru, Machiel de Jager, J.H. Fred Falkenburg, Jack A.M. Leunissen & Peter A. van Veelen.

Supplemental material is available at:
http://link.springer.com/article/10.1007%2Fs00251-010-0497-1

Harm Nijveen & Michel Kester share first authorship and Jack Leunissen & Peter van Veelen share senior authorship

ABSTRACT

T cell epitopes derived from polymorphic proteins or from proteins encoded by alternative reading frames (ARFs) play an important role in (tumour) immunology. Identification of these peptides is successfully performed with mass spectrometry. In a mass spectrometry-based approach the recorded tandem mass spectra are matched against hypothetical spectra generated from known protein sequence databases. Commonly used protein databases contain a minimal level of redundancy and thus are not suitable data sources for searching polymorphic T cell epitopes, either in normal or ARFs. At the same time, however, these databases contain much non-polymorphic sequence information, thereby complicating the matching of recorded and theoretical spectra, and increasing the potential for finding false positives. Therefore, we created a database with peptides from ARFs and peptide variation arising from single nucleotide polymorphisms (SNPs). It is based on the human mRNA sequences from the well-annotated reference sequence (RefSeq) database and associated variation information derived from the Single Nucleotide Polymorphism Database (dbSNP). In this process we removed all non-polymorphic information. Investigation of the frequency of SNPs in dbSNP revealed that many SNPs are non-polymorphic "SNPs". Therefore, we removed those from our dedicated database, and this resulted in a comprehensive high quality database, which we coined the Human Short Peptide Variation database (HSPVdb). The value of our HSPVdb is shown by identification of the majority of published polymorphic SNP- and/or ARF-derived epitopes from a mass spectrometry-based proteomics workflow, and by a large variety of polymorphic peptides identified as potential T cell epitopes in the HLA-ligandome presented by Epstein-Barr virus cells.

INTRODUCTION

T cell-mediated immunotherapy is an attractive treatment of cancer as it exploits the potential of cytolytic T cells to specifically recognize antigens that are selectively expressed on tumour cells (Storb 2003; Hambach and Goulmy 2005; Kessler and Melief 2007; Falkenburg et al. 2003; Bleakley and Riddell 2004; Eisenlohr 2007). The enormous specificity of T cells involved in killing tumour cells makes this kind of treatment very attractive. An excellent example is the powerful graft-versus-leukaemia (GVL) effect witnessed after allogeneic hematopoietic stem cell transplantation. GVL is characterized by remission of a haematological malignancy coinciding with the *in vivo* expansion of tumour-specific T cells. These T cells react to a patient-specific epitope presented in human leukocyte antigen (HLA) molecules on tumour cells (Marijt et al. 2003; van Bergen et al. 2007). T cell epitopes are peptides with a length of generally 8-11 amino acids. T cells are capable of distinguishing epitopes differing by only one amino acid, caused by a single nucleotide difference between patient and donor (Spierings et al. 2007). T cell epitopes, identified to play a role in (tumour) immunology, may arise from regular reading frames, but can also be encoded by alternative reading frames (ARFs) (Ho et al. 2006). Given the need for therapeutically useful T cell epitopes, the identification of new epitopes is of unceasing importance. The identification of T cell epitopes has been achieved with an array of methods, among which mass

spectrometry is one of the most prominent techniques (Engelhard 2007; Hillen and Stevanovic 2006; Nesvizhskii et al. 2007). Peptide identification by tandem mass spectrometry is most successfully applied in an ever increasing number of proteomics studies. In a typical high throughput proteomics/ligandomics setting (Oliveira et al. 2010), the experimentally determined tandem mass spectra are matched against a database of hypothetical spectra generated from known peptide sequences using search engines like Mascot (Perkins et al. 1999) and Sequest (Eng et al. 1994).

For mass spectrometry-based identification of epitopes from polymorphic proteins, like minor histocompatibility antigens (MiHA) and peptides arising from ARFs, the commonly used protein databases like UniProt (The UniProt Consortium 2008), IPI (Kersey et al. 2004) and RefSeqP (Pruitt et al. 2007) are unsuitable data sources, since these display very incomplete information about polymorphisms. Most of the published polymorphic MiHA are, therefore, not present in the standard protein databases, used in mass spectrometry-based workflows. Several strategies have been employed to address this problem (MSIPI (Schandorff et al. 2007), PepHum (Edwards 2007)), each with its own merits and limitations, trying to find the right balance between database size and completeness. In addition, there is a wealth of ligand and/or epitope information databases (Salimi et al. 2010), but these are not applicable in mass spectrometry (MS)-based workflows. Knowing that customized search databases that provide detailed control over the search space can vastly outperform standard strategies (Reisinger and Martens 2009), we designed a database dedicated to MiHA, thereby improving the chance of their identification in a proteomics type of experimental set up.

Our approach is based on the coding potential of the human genome, including its documented variations, as described in the RefSeq database. We chose RefSeq because it contains minimal redundancy, while still retaining splice variants, incorporates single nucleotide polymorphism (SNP) data from Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al. 2001), which are richly annotated. We have created a database that contains all possible short peptides in different reading frames from a non-redundant mRNA set, combined with the known and annotated variations/SNPs. In this process we removed all non-polymorphic information. Investigation of the frequency of SNPs in dbSNP revealed that many of these SNPs are non-polymorphic "SNPs". Therefore, we removed those from our dedicated database as well, and this resulted in a high quality comprehensive polymorphic peptide database. Centred on the amino acid polymorphisms of non-synonymous SNPs, our dedicated Human Short Peptide Variation database (HSPVdb) outperforms existing databases in MS/MS-based T cell epitope identification.

The value of our HSPV database is shown by identification of the majority

of published polymorphic SNP- and/or ARF-derived epitopes from a mass spectrometry-based proteomics workflow, as well as by a large variety of polymorphic peptides identified as potential T cell epitopes in the HLA-ligandome presented by EBV cells.

MATERIALS AND METHODS

Database preparation

The HSPVdb consists of peptides derived from genomic sequence variations. The database only contains peptides of seven amino acids or longer. The RefSeq database release 32 was downloaded from the NCBI FTP site and indexed using our local SRS installation (Etzold et al. 1996), (http://srs.bioinformatics.nl). The human mRNA subsection of RefSeq was extracted by selecting records with molecule type "mRNA" and organism source "Homo sapiens". The resulting list of RefSeq records was subsequently processed using a series of Perl scripts.

To create the peptides derived from genomic sequence variations, we made use of the variation annotations that were added to RefSeq by the dbSNP staff. Variations found in the 5' and 3' UTRs were purposely included to allow detection of T cell epitopes derived from ARFs. For each annotated variation, the nucleotide sequences corresponding to the different alleles were generated. Instead of duplicating the complete mRNA sequence for each allele, we took a fragment starting 30 nucleotides upstream and ending 32 nucleotides downstream of the variation. The three forward reading frames of each allele were translated to amino acid sequences. This typically results in three peptide sequences of 20 amino acids. Translation ignored the presence or absence of start codons. Codons that could not be translated to a single amino acid due to ambiguous nucleotides were translated to a stop codon. The amino acid translation was split on stop codons to get peptides derived from a continuous reading frame. Only the peptides including the variation were kept in the database. To minimize redundancy, a translation for an allele was only included when the variation gives rise to a change in amino acid sequence (non-synonymous SNPs). This part of the database is optimized for finding peptides in the size range between 8 and 11 amino acids, but databases containing other peptide lengths can be produced at will. The database presented here consists of 20-mer peptides.

Each peptide sequence that was created was stored as a separate database record and annotated with the ID of the originating mRNA sequence and the location of its encoding reading frame. If the RefSeq entry contains a coding sequence (CDS), the protein identifier and the position of that CDS on the mRNA with corresponding protein identifier were added as annotation to the database record. For variations, we included the corresponding dbSNP identifiers, the positions of the variations, the nature of the amino acid changes and the percentage

heterozygosity. If a variation causes an amino acid substitution, a SAP (single amino acid polymorphism), the possible amino acids were listed. Insertions or deletions were annotated as "in/del". The resulting database was stored as a flat file in FASTA format for mass spectrometry-based proteomics purposes. This HSPVdb is fully dedicated to finding polymorphic epitopes. To reduce the size of this database, all duplicate amino acid sequences were deleted. These peptides contain both polymorphisms for each position, thereby describing all possible SNP information.

Subsets of the HSPV database were created based on reported heterozygosity. Three heterozygosity categories were defined: 0/1, unknown, all others. Additionally, for all categories ARFs were either included or left out.

Peptides for which the encoding DNA sequence is not part of the in RefSeq-annotated open reading frame are labelled as alternative reading frame or ARF peptides. These include CDS that are in a different reading frame and sequences that are located up- or downstream of the annotated open reading frame.

SNP genotyping assays

Genomic DNA was isolated from 192 HLA A*0201 positive patient and donor samples (peripheral blood mononuclear or bone marrow cells) by the Gentra Systems PUREGENE genomic isolation kit (Biocompare, San Francisco, CA). SNPs rs4848158, rs61378134, rs36023150, rs11540526, rs11554279, rs35958189, rs56013141, rs11541290, rs34422048, rs11541416, rs28659989, rs2070159, rs4261080, rs11557142, rs11555631, rs11479605, rs11541519, rs5030742, rs11548263 were analysed using a KASpar assay with allele-specific primers labelled with VIC and FAM dyes, (KBioScience, Hoddesdon, UK). Genotyping was performed according to manufacturer's instructions.
Illumina Custom Array was used for genotyping rs10960, rs1143138, rs12986002, rs34669146, rs1047844, rs11266765, rs11539866, rs11541416, rs11541519, rs11542419, rs11542836, rs11544489, rs11545551, rs11548082, rs11553285, rs11553982, rs11554156, rs11554279, rs11555631, rs11557142, rs11558570, rs13202878, rs17848351, rs17851857, rs17853301, rs17853718, rs1803181, rs2070159, rs2261324, rs28934887, rs28935171, rs28940302, rs3180961, rs34136999, rs34418712, rs3962697, rs4848158, rs5030742, rs6112008, rs6686209, rs6794514
Genotyping was performed according to manufacturer's instructions.

Sample preparation for test set

Peptide synthesis

Peptides were synthesized by standard Fmoc chemistry on a Syro II peptide synthesizer as described previously (Hiemstra et al. 1997). The integrity of the

Epitope name* / HLA	Sequence	Remarks	Gene	polymorphic AA	dnSNP entry
HA1 / A2	VLHDDLLEA	immunogenic	HMHA1	VL[R/H]DDLLEA	rs1801284
HA2 / A2	YIGEVLVSV	immunogenic	MYO1G	YIGEVLVS[V/M]	rs61739531
HA3 / A1	VTEPGTAQY	immunogenic	AKAP13	V[M/T]EPGTAQY	rs2061821
HA8 / A2	RTLDKVLEV	immunogenic	KIAA0020	[R/P]TLDKVLE[V/I]	rs2270891
HA1 / B60	KECVLHDDL	immunogenic	HMHA1	KECVL[R/H]DDL	rs1801284
LB-ADIR-1F / A2	SVAPALALFPA	immunogenic; ARF in 5'UTR	TOR3A (ADIR)	SVAPALAL[F/S]PA	rs2296377
LB-ADIR-1S / A2	SVAPALALSPA	allelic counterpart			
CTSHr / A31	ATLPLLCAR	immunogenic	CTSH	ATLPLLCA[G/R]	rs2289702
CTSHr / A33	WATLPLLCAR	immunogenic	CTSH	WATLPLLCA[G/R]	rs2289702
ACC1y / A24	DYLQYVLQI	immunogenic	BCL2A1	DYLQ[C/Y]VLQI	rs1138357
ACC1c / A24	DYLQCVLQI	immunogenic			
ACC1c +cystinylated	DYLQCVLQI	immunogenic			
HB1h / B44	EEKRGSLHVW	immunogenic	HMHB1	EEKRGSL[H/Y]VW	rs161557
ACC2d / B44	KEFEDDIINW	immunogenic	BCL2A1	KEFED[G/D]IINW	rs3826007
ACC2g / B44	KEFEDGIINW	allelic counterpart			
LB-ECGF1-1H/B7#	RPHAIRRPLAL	immunogenic; ARF	TYMP(ECGF1)	RP[H/R]AI[R/C]RPLAL	no entry; rs1061205

Table 1. Overview of known MiHA, used as a test set in this study. It displays the epitope name and the HLA-molecule it is presented in. In addition, its immunogenicity is indicated together with the gene name and the polymorphisms are indicated.

*Names according to: http://www.lumc.nl/dbminor.

peptides was checked by reversed-phase high-performance liquid chromatography (HPLC) and mass spectrometry.

Liquid chromatography-mass spectrometry

The peptides studied are listed in Table 1. These are minor histocompatibility antigens as identified by different research groups around the world. A more complete listing of MiHA can be found at http://www.lumc.nl/dbminor. To perfectly mimic the conditions used in a normal mass spectrometry-based HLA-ligand identification process, all peptides included in Table 1 were measured by on-line chromatography/mass spectrometry (see below), and tandem mass spectra were recorded of their singly, doubly and triply charged form. Subsequently, a selection of relevant charge states was made for each peptide, and charge states with a substantial contribution to the overall intensity only were used to construct a Mascot generic file (MGF) containing 31 tandem mass spectra, see Table 2.

Sample preparation for determination of the EBV-LCL ligandome

Cell collection, preparation & HLA elutions

Peripheral blood samples were obtained from healthy donors after approval by the Leiden University Medical Center Institutional Review Board and informed consent according to the Declaration of Helsinki. Mononuclear cells (MNC) were isolated by Ficoll-Isopaque separation and cryopreserved. Stable Epstein-Barr virus (EBV)-transformed B cell lines (EBV-LCL) were generated using standard procedures. EBV-LCL and HeLa cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM, BioWhittaker, Verviers, Belgium) supplemented with 10% bovine fetal serum (FBS, BioWhittaker).

Peptide isolation

Peptide isolation was performed with protein A beads (GE healthcare) covalently linked to the major histocompatibility complex (MHC) class I mAb W6/32 (3 mg W6/32 on 1 ml of ProtA sepharose) using dimethyl pimelimidate according to the standard protocol (Stepniak et al 2008).

The complex MHC-peptide pool was prefractionated on a C18 RP-HPLC system (2 mm x 15 cm; Reprosil-C18-AQ 3 um; Dr. Maisch GmbH, Ammerbuch, Germany), using a gradient 0-60% A to B. A: water, 5% Acetonitrile (ACN), 0.1% TFA, B: ACN, 0.1% TFA

Liquid chromatography-mass spectrometry

Peptide fractions were reduced to near dryness and resuspended in 95/3/0.1 v/v/v water/acetonitrile/formic acid. These resuspended fractions were analysed by on-

line nano-HPLC mass spectrometry with a system described by Meiring et al (Meiring et al. 2002). Fractions were injected onto a precolumn (100 um x 15 mm; Reprosil-Pur C18-AQ 3 um, 5 um, Phenomenex) and eluted via an analytical nano-HPLC column (15 cm x 50 um; Reprosil-Pur C18-AQ 3 um). The gradient was run from 0% to 50% solvent B (10/90/0.1 v/v/v water/acetonitrile/formic acid) in 90 min. The nano-HPLC column was drawn to a tip of approximately 5 um and acted as the electrospray needle of the MS source.

The mass spectrometer was an LTQ-FT Ultra (Thermo, Bremen, Germany) and was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition. Full scan mass spectra were acquired in the FT-ICR with a resolution of 25,000 at a target value of 5,000,000. The two most intense ions were then isolated for accurate mass measurements by a selected ion monitoring scan in FT-ICR with a resolution of 50,000 at a target accumulation value of 50,000. The selected ions were then fragmented in the linear ion trap using collision-induced dissociation at a target value of 10,000. In a post analysis process, raw data were converted to peak lists using Bioworks Browser software, Version 3.1. For peptide identification, MS/MS data were submitted to the human IPI database using Mascot Version 2.2.04 (Matrix Science) with the following settings: 2 ppm and 0.8-Da deviation for precursor and fragment masses, respectively; no enzyme was specified. The Mascot output files were loaded into Scaffold (http://www.proteomesoftware.com) and exported to Excel as peptide reports and duplicates were removed.

RESULTS

To investigate the value of our database we studied two sets of samples. First, a test set comprising approximately 30% of all MiHA known today, as listed in Table 1, and second, a set of peptides eluted from HLA from an EBV-cell line.

Validation of HSPVdb with a test set of known MiHA

Our test set of known polymorphic peptides and allelic counterparts were synthesized, and measured in standard on-line nano-HPLC/MS experiments, as in our normal proteomics workflow on HLA-ligands (Oliveira et al. 2010). Of all significantly occurring charge states tandem spectra were recorded. Tandem mass spectra of varying quality are present in this dataset, reflecting a “real-world” situation, where the spectral quality depends on intrinsic peptide properties. A combined peak list was constructed from these spectra for searching the databases used in this work. This led to a set of 31 experimental tandem MS derived from 15 peptides (Table 2).

For validation of our HSPVdb, we compared it to the MSIPI and PepHum databases that were specifically constructed to address the lack of peptide variation in common databases like IPI. A summary of the databases used in this study is shown in Table 3.

			Database	IPI369	
			Mass accuracy (ppm)	1	
			Mascot cut-off score	37	37
Peptide name*	**Sequence**	**Charge**	**Pr?**	**sco**	**int**
CTSHr A31	ATLPLLCAR	2		10	wr
CTSHr A33	ATLPLLCAR	1			np
HA3t A1	VTEPGTAQY	2		12	wr
HA3t A1	VTEPGTAQY	1		9	wr
HA2v A2	YIGEVLVSV	1	Y	18	wr
LB-ADIR-1S A2	SVAPALALSPA	2		22	wr
LB-ADIR-1S A2	SVAPALALSPA	1		13	wr
HA1h A2	VLHDDLLEA	2		28	wr
HA1h A2	VLHDDLLEA	1		26	wr
LB-ADIR-1F A2	SVAPALALFPA	2		17	wr
LB-ADIR-1F A2	SVAPALALFPA	1		25	wr
HA1h B60	KECVLHDDL	2		14	wr
HA1h B60	KECVLHDDL	1		5	wr
HA8rv A2	RTLDKVLEV	3	Y	37	Y
HA8rv A2	RTLDKVLEV	2	Y	34	wr
HA8rv A2	RTLDKVLEV	1	Y	32	ye
ACC1c	DYLQCVLQI	2	Y	50	Y
ACC1c	DYLQCVLQI	1	Y	36	wr
CTSHr A33	WATLPLLCAR	2		8	wr
ACC1y BCL2A1-A24	DYLQYVLQI	2		27	wr
ACC1y BCL2A1-A24	DYLQYVLQI	1		22	wr
ACC1c+cys	DYLQCVLQI	2	Y	42	Y
ACC1c+cys	DYLQCVLQI	1	Y	36	ye
HB1h B44	EEKRGSLHVW	3	Y	10	wr
HB1h B44	EEKRGSLHVW	2	Y	16	ye
ACC2g BCL2A1-B44	KEFEDGIINW	2	Y	48	Y
ACC2g BCL2A1-B44	KEFEDGIINW	1	Y	34	ye
LB-ECGF-1H B7#	RPHAIRRPLAL	3		5	wr
LB-ECGF-1H B7	RPHAIRRPLAL	2			np
ACC2d BCL2A1-B44	KEFEDDIINW	2		17	wr
ACC2d BCL2A1-B44	KEFEDDIINW	1		15	wr

Table 2. Summary of the searches with the test set of known MiHA against the IPI, MSIPI, PepHum and HSPV databases. The peptide names and sequences are given

MSIPI367			PepHum			HSPVdb		
	1			1			1	
	38	38		44	44		28	28
Pr?	**sco**	**int**	**Pr?**	**sco**	**int**	**Pr?**	**sco**	**int**
Y	42	Y	Y	42	ye	Y	42	Y
Y		np	Y	8	ye	Y		np
Y	34	ye	Y	34	ye	Y	34	Y
Y	18	ye	Y	18	wr	Y	18	ye
Y	18	wr	Y	28	wr	Y	16	ye
	22	wr	Y	48	Y	Y	48	Y
	13	wr	Y	34	wr	Y	8	wr
Y	28	wr	Y	28	wr	Y	15	ye
Y	40	Y	Y	40	ye	Y	40	Y
	17	wr		28	wr	Y	66	Y
	25	wr		29	wr	Y	4	wr
Y	36	ye	Y	36	ye	Y	36	Y
Y	29	ye	Y	29	ye	Y	29	Y
Y	37	ye	Y	37	ye	Y	37	Y
Y	34	wr	Y	34	wr	Y	30	Y
Y	32	ye	Y	32	ye	Y	32	Y
Y	50	Y	Y	50	Y	Y	50	Y
Y	36	wr	Y	40	wr	Y	15	ye
Y	37	ye	Y	37	ye	Y	37	Y
Y	58	Y	Y	58	Y	Y	58	Y
Y	25	ye	Y	25	ye	Y	25	ye
Y	42	Y	Y	42	ye	Y	42	Y
Y	36	ye	Y	36	ye	Y	36	Y
Y	10	wr	Y	13	wr	Y	6	ye
Y	16	ye	Y	16	wr	Y	16	ye
Y	48	Y	Y	48	Y	Y	48	Y
Y	34	ye	Y	34	ye	Y	34	Y
	5	wr		16	wr		3	wr
		np		4	wr			np
Y	39	Y	Y	39	ye	Y	39	Y
Y	45	Y	Y	45	Y	Y	45	Y

together with the charge of the precursor, submitted to tandem mass spectrometry.
(continued on next page) *Names according to: http://www.lumc.nl/dbminor.

Table 2 (cont). For each database three columns are displayed: (1) whether the peptide is present in the database (*Pr?*), followed by (2) the mascot ion score assigned to the tandem mass spectrum and (*black filling* if the mascot ion score is above the threshold of the search) (3) the evaluation, i.e. was the tandem mass spectrum matched to the correct peptide (*black filling* and (*Y*) if correct and above the mascot threshold (cut-off score), *grey filling* if correct and below (*ye*) the mascot threshold. In short, the blacker, the better. The HSPVdb scores very well, due to its reduced format in combination with a high density of relevant SNP information. *Wr* wrong interpretation of MS2 spectrum; *np* no matching/no proposal from mascot search. #Charge state 4+ was the most abundant in the charge distribution of peptide LB-ECGF-1H, but its MS2 spectrum was of such poor quality that it was not included for database searching. LB-ADIR peptides are from an ARF. ACC1+ Cys represents a special case in which the cysteine residue in the epitope can be modified by formation of an S-S bridge with free cysteines. This is relevant for both *in vivo* recognition and mass spectrometric interpretation.

The HSPVdb is similar to the size of the IPI and MSIPI databases, but it includes all SNP information in all forward and alternative reading frames (MSIPI: 170.242 SNPs; HSPVdb: 380.182 SNPs). When leaving out the alternative reading frame information (i.e., HSPVdb subset 1, see Table 3) the size of our HSPVdb is reduced to only 25% of the size of IPI and MSIPI, which is of great importance when searching databases.

The test set containing the tandem mass spectra of known MiHA was searched against the IPI, MSIPI, PepHum and our HSPVdb. Searches were performed, using the Mascot search engine (Matrix science), with various settings for mass accuracy (1, 2, 5, 10 and 50 ppm) representing the mass accuracy of various MS and/or experimental set-ups. The enzyme setting was “none”. It is important to note that in the elucidation of HLA-ligands the peptide termini are unknown in contrast to the vast majority of cases in standard proteomics experiments, in which peptide matching against databases can be done with an additional and very stringent condition, namely an enzyme cleavage site (in most cases trypsin). In the standard proteomics approach the enzyme restriction has an enormous positive impact on specificity and search time. For the sequencing of T cell epitopes, enzyme restriction is not applicable. However, for binding to the presenting HLA molecule, HLA-ligands have to satisfy certain conditions imposed by the HLA molecule, the binding motif. This binding motif can be used as additional help to some extent to assess the value of the matched sequence by the search engine.In addition, netMHC, http://www.cbs.dtu.dk/services/NetMHC/, could be applied to some extent, but neither of the two can be directly applied in the database search as a fixed condition. The best proof of a correct peptide assignment, in spite of improvements in peptide matching algorithms, is still the comparison of the tandem spectrum of the proposed eluted epitope with its synthetic counterpart.

Database	Number of sequences	Number of residues	Size relative to IPI 3.69	ARFs?	0/1?	Unk?
IPI (HUMAN v3.69)	87130	35200044	1.00		n.a.	n.a.
MSIPI (HUMAN v3.67)	87040	42553286	1.21		✓	✓
PepHum	75237	176019757	5.00	✓	✓	✓
HSPV	2634086	45422884	1.29	✓	✓	✓
			Rel. to set 5			
HSPV subset 1	423015	8344552	0.18		✓	✓
HSPV subset 2	377269	7440614	0.16			✓
HSPV subset 3	106379	2108989	0.05			
HSPV subset 4	152125	3012927	0.07		✓	
HSPV subset 5	2634086	45422884	1.00	✓	✓	✓
HSPV subset 6	2378073	41106669	0.90	✓		✓
HSPV subset 7	729721	12444311	0.27	✓		
HSPV subset 8	985734	16760526	0.37	✓	✓	

Table 3. Overview of the databases used in this study, listing the number of entries and the number of amino acid residues present in each database. In addition, the presence of ARFs and the (type of) SNP information in the various databases is indicated. The number of residues of each database relative to the IPI database and the relative size of the HSPV subsets is given. The number of SNPs in MSIPI 3.67 is 170.242; the number of SNPs in HSPVdb (subsets 1 en 5) is 380.182.

All output of the Mascot search engine was assessed manually, and a summary of the results for a 1-ppm mass accuracy is shown in Table 2, and a full report of the searches is given in Supplementary Table 1.

Table 2 shows a selection of the searches in the four databases with a 1-ppm mass measurement accuracy. For every individual tandem mass spectrum the Mascot ion score is reported. The results from the database search were classified by the following criteria: (1) was the tandem mass spectrum correctly identified by the search engine (indicated by black and grey filling in the first column for each database)? and (2) was the identification score above (indicated by black filling in the second column for each database) or below the Mascot significance threshold (cut-off score)? Therefore, "the blacker the better". The presence ("Pr") of each peptide in the particular database is indicated by "Y" in the appropriate column. Supplementary Table 1 shows the results of all searches performed with the test set of 31 tandem mass spectra to the IPI 3.69, MSIPI 3.67, PepHum and HSPVdb. From Table 2, it is immediately clear that the IPI database is not useful for finding MiHA, since it lacks essential variation information.
The PepHum database, based on expressed sequence tags (ESTs) information, including ARFs, is relatively large, by which relevant information for finding our polymorphic epitopes is "diluted" and consequently a serious amount of "noise" is generated, increasing the chance of finding false positives. The consequence of this is reflected in the outcome of the database search for PepHum. The number of significantly scoring peptides is only 5 as compared to the 19 peptides identified by our HSPVdb, see also Figure 1a. This low score is only partially rescued by the number of correctly assigned peptides with a score below the Mascot significance threshold. In addition, ESTs may be more prone to experimental sequencing errors, leading to occurrence of false SNPs.
The elegantly produced MSIPI does quite well, but also here, most correct peptide hits are below the statistical significance threshold score, which makes it hard to decide if a hit is true or a false positive in a "non-test set" setting. In addition, the MSIPI does not contain information from ARFs and UTRs.
For the HSPVdb, out of 31 MS/MS spectra, 19 are identified correctly above the Mascot significance threshold, while another 7 are also correctly identified, but below the significance threshold. Only three tandem mass spectra were wrongly assigned (false positives). These wrong assignments are caused by the poor quality of the tandem mass spectra of these peptides, due to intrinsic peptide properties. To two tandem mass spectra no match was assigned. These tandem mass spectra represent two peptides, "YIGEVLVSV", which yields a bad mass spectrum and "RPHAIRRPLAL", which is not present in the HSPVdb subset because, it is derived from a SNP not found in the dbSNP database. The HSPVdb, designed to reduce non-informative sequence information, outperforms the other databases.
Next to the size of the database, relieving the accuracy condition from 1 to 50

ppm (Figure 1b) has a detrimental effect on both the number of correctly assigned peptides above and below the Mascot significance threshold. This effect can even lead to a false-positive score, as illustrated by a high and significant Mascot score of 63 (!) for MS/MS spectrum/query #6 (in HSPVdb, 50 ppm), see supplementary Table 1a. This result emphasizes the value of high mass accuracy.

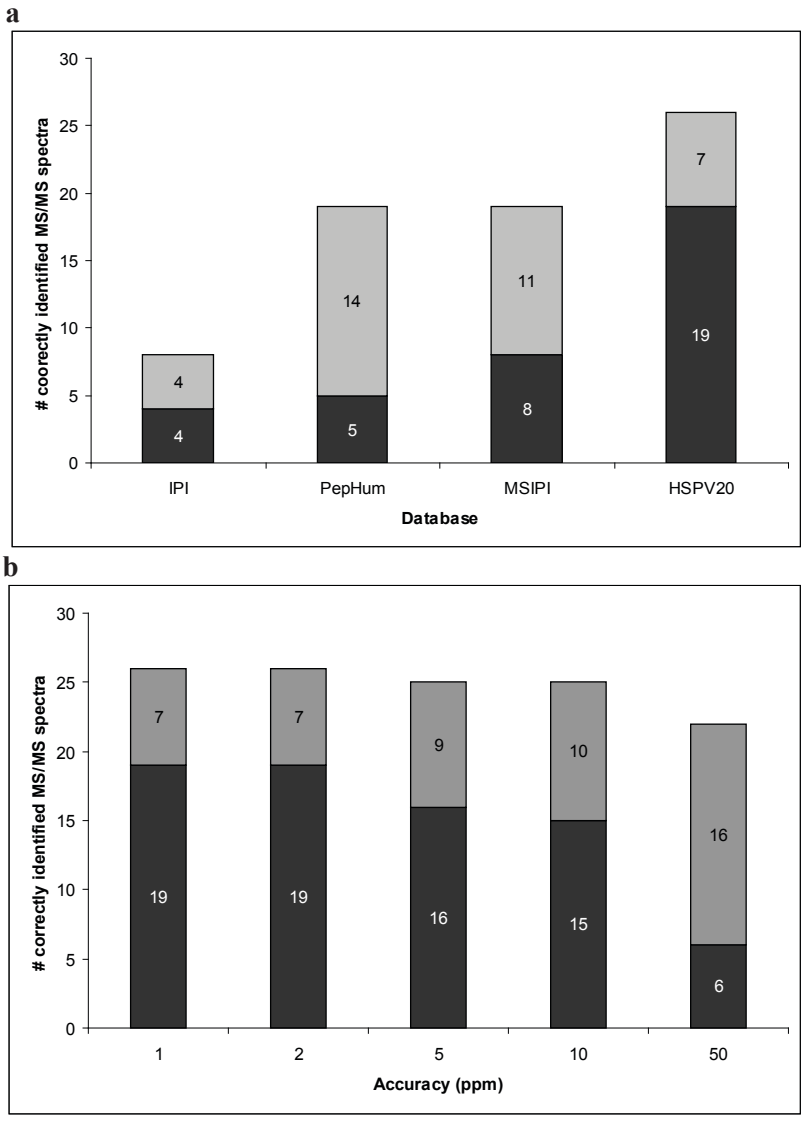


Figure 1. a. Summary of the searches with 1-ppm accuracy against the IPI, MSIPI, PepHum and HSPV databases. The colour coding is as follows: *black* correct hit and above the MASCOT significance threshold; *grey* correct hit, but below the significance threshold. **b.** Summary of the searches against HSPVdb with various mass measurement accuracies, 1, 2, 5, 10 and 50 ppm accuracy against the HSPV database. The colour coding is as above.

So far, the good performance in the MS/MS-based identification of T cell epitopes of HSPVdb can be attributed to the compact nature and the special focus on polymorphic peptides. A reduced database size directly translates to a lower noise level in the database search, which is especially important in high throughput T cell epitope elucidation, where search space limiting constraints like an enzyme cleavage site cannot be used. Another parameter affecting search quality is mass accuracy, which is also proven to be a prominent factor in avoiding false positives. To further improve the quality of our HSPVdb, we focused on the quality of the SNPs in dbSNP, since we noted that the reported frequency of a substantial number of SNPs in dbSNP is "0" or "1" or "unknown". This made us decide to study a random set of 52 SNPs with no frequency reported in dbSNP. We developed a SNP assay to screen a random HLA-A*02-positive Dutch donor population using the KASPar assay (92 DNA samples) and a SNP array (192 DNA samples). In our test population, 46 out of the 52 SNPs (90%) were not polymorphic, having an allele frequency of 1 or 0 in the SNP assays. Two SNPs (4%) were very rare (allele frequencies of 0.97, and 0.99), and 4 SNPs (8%) had a reasonable distribution in our population (0.77; 0.70; 0.20; 0.13).

A large number of reported "SNPs" in dbSNP is apparently not polymorphic, thereby contaminating our proteomics approach and the chance of finding suitable patient/donor MiHA pairs. Therefore, since reduction of the search space greatly enhances the chance of finding true positives in database searches, we decided to test our HSPVdb after removal of either "unknowns" or "0" and "1", or both. The results are shown in supplementary Table 1b. Subset 3, the leanest form of HSPVdb with both "0" and "1" and "unknown" frequency' SNPs removed and without ARFs, is reduced to only one fourth of its original size. Therefore, the significance threshold is clearly lowered (from 28 to 22 for 1-ppm mass accuracy), increasing the chance of finding true positives. In particular, those derived from tandem mass spectra of relatively poor quality with accompanying intrinsic low Mascot scores. Only one true positive is lost, because its frequency is not reported in dbSNP. Similarly, the other subsets (subsets 1-8) of HSPVdb have reduced significance thresholds (data not shown). The application of these various forms of the database can be adapted to the needs of the user.

So far, we have shown that selective reduction of the database size by removal of both the non-polymorphic peptide stretches and the SNPs of limited value, leads to a comprehensive high quality database file dedicated to improving the elucidation of MiHA.

Database quality and inconsistencies

During this work we discovered inconsistencies in the number of SNPs included in several RefSeq and MSIPI versions, see Figure 2a and 2b.

The number of reported human SNPs dropped by 50% going from RefSeq release 28 to release 30, and by more than 50% in MSIPI going from version 37 to version 38. We reported this in October 2008 to the respective database producers who acknowledged there were problems and improved their efforts. Recently we encountered a problem with the SNPs reported by 1000genomes.org in dbSNP, which is being solved. Therefore, we continued using version 3.32 (on our website the HSPVdb version based on either Refseq release 32 or release 40 can be chosen). We would like to warn users for the status of the RefSeq with

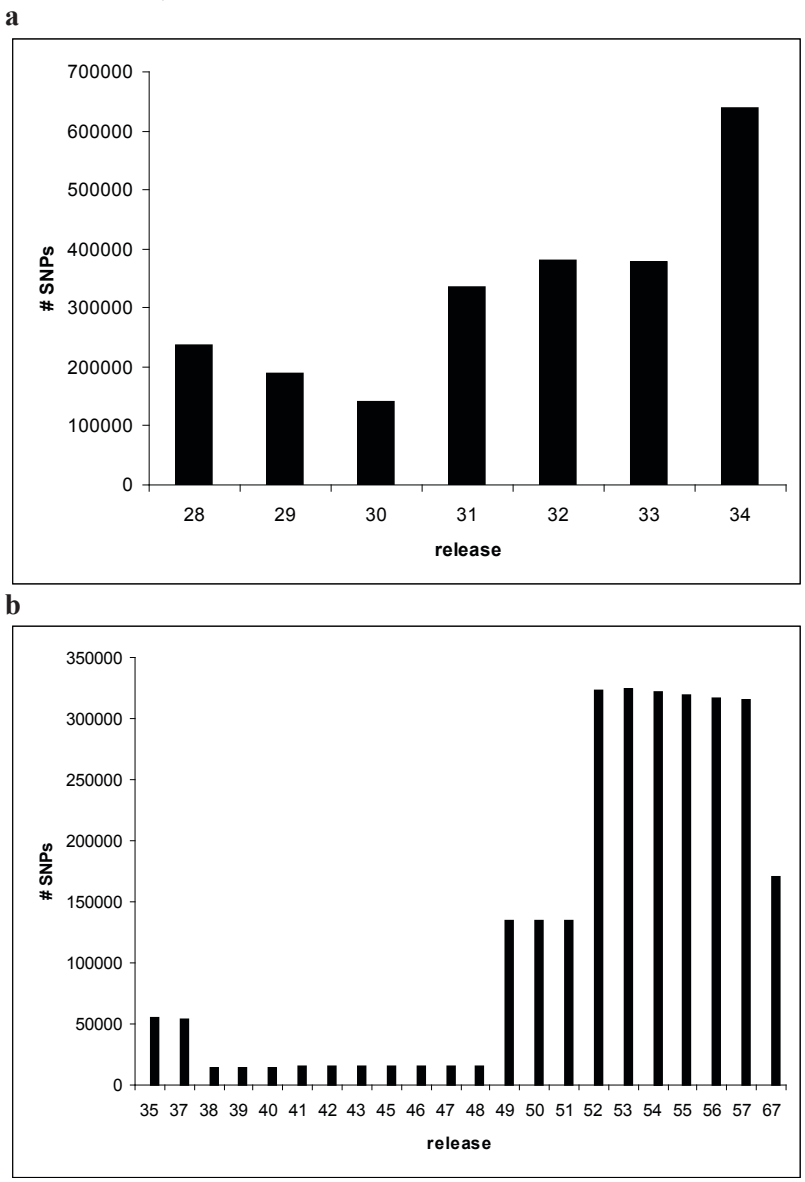


Figure 2. Number of incorperated SNPs per release of RefSeq (**a**) and of MSIPI (**b**)

respect to this. MSIPI, also being a secondary database, suffered from the same errors during several versions, but this has been repaired, starting from version 49, although a strong decrease can be seen in version 3.67 (Figure 2b). In general, as a user of these databases, it is very hard to judge the value of the databases, so caution should be taken: newer versions are not always better.

Application of HSPVdb to finding potential MiHA presented in HLA on EBV-cells

To investigate the effects of application of our database to a representative HLA-ligand elution experiment, we eluted peptides from an EBV-LCL cell line (EBV-JY). After lysis, affinity purification was performed with BB7.2 antibody for HLA-A2, followed by separation of HLA and peptides. Subsequently, the complex peptide pool was analysed by on-line nano-HPLC-tandem MS. The tandem mass spectra were matched against several databases for comparison, in particular MSIPI and various subsets of our HSPV database.
Here, MSIPI is compared to the smallest subset of our HSPV database without ARFs (subset 3) and with ARFs included (subset 7), the advantages of which have been illustrated for the test set described above. These trimmed subsets do not include SNPs of which the frequencies in dbSNP are reported to be 0/1 or unknown. By searching against the smaller compact database containing all relevant SNPs, intermediate scoring peptides appear in the database search that would otherwise fall below the significance threshold when matching tandem mass spectra against much larger databases.

This is illustrated by the number of intermediate scoring peptides, i.e. those peptides that score below the Mascot significance threshold when matching against MSIPI, and are, therefore, peptides not found otherwise. An additional 130 peptides were found for subset 3 and an additional 400 for subset 7. These extra peptides need to be checked for false positives (peptides with tandem mass spectra that match better with non-SNP containing peptides), and for the presence of a SNP. The extra peptides found can e.g. be evaluated by application of netMHC. This approach, starting from our small experimental elution experiment, yielded eight peptides from subset 7 (including ARFs), and five peptides from subset 3 with a netMHC score below 50 (i.e., a stringent condition for strong binding). These peptides, shown in Table 4, are currently evaluated as potential MiHA.

All peptides found only by searching against the dedicated HSPV database increase the chance of finding relevant MiHA. The excellent annotation of the SNPs reported in our HSPV database enables the user to directly jump to the relevant information about the polymorphism, a feature that was largely lacking so far.
The HSPV database described here is an integral part of a complete peptidomics pipeline for finding therapeutically useful MiHA, a strategy that is currently

under development.

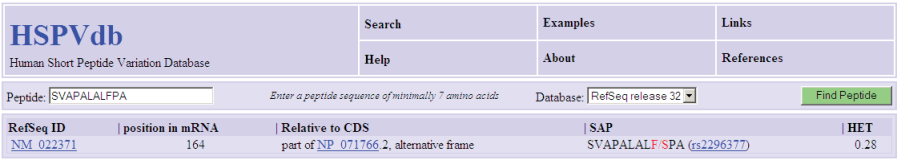
Availability and Web interface

A flat file with the content of the HSPV database can be requested by sending an email to hspv@bioinformatics.nl. A simple interactive query interface is available at: http://srs.bioinformatics.nl/hspv/

This web interface allows the biologist to query the database for peptide sequences. It returns a list of RefSeq mRNA entries that contain a continuous reading frame encoding the query peptide, the start position of that reading frame, the position of the encoding nucleotide sequence with respect to any annotated CDS, and a description of the variations if the peptide contains any, see Figure 3a. This is a great feature for the initial assessment of the quality and potential usefulness of the output of our database searches.

The richness of SNP information of our database is shown in Figure 3b, for the peptide "TLSELHCD" displaying SAPs at every position in the peptide.

a



b

Figure 3. Screen shots show the output of a query for the peptides SVAPALALFPA (**a**) and TLSELHCD (**b**). It clearly illustrates the effect of the large number of annotated variations at the amino acid level.

Peptide	mRNA	Gene	Protein	rel2cds	in frame	dbSNP	SNP	Het	NetMHC
FLIPKTLVGV	NM_017700	FLJ20184	NP_060170.1	downstream	y	rs2121558	FLIPKTLVG[E/V]	0.47	9
SLSDLIYAL	NM_001080837	SEBOX	NP_001074306.2	inside	y	rs9910163	SLSDLIYA[L/S]	0.13	7
GLWEQENHL	NM_024713	C15orf29	NP_078989.1	inside	y	rs34998154	GLW[E/K]QENHL	0.05	41
FIVTVIHTI	NM_024607	PPP1R3B	NP_078883.2	downstream	n	rs330915	FIVTVIHT[I/F]	0.49	30
FLSEHPNVTL	NM_145298	APOBEC3F	NP_660341.2	inside	y	rs17000697	FL[A/S]EHPNVTL	0.28	19
FLNQRSIML	NM_030956	TLR10	NP_112218.2	upstream	n	rs9998678	FLNQ[R/W]SIML	0.05	29
LLQSLVSI	NM_198889	ANKRD17	NP_942592.1	inside	n	rs6855349	LLQS[S/L]VSI	0.46	46
TLLDPNEKYLL	NM_016243	CYB5R1	NP_057327.2	inside	y	rs2232842	TLLDP[N/S]EKYLL	0.31	31

Table 4. Exclusive peptides with selected info from the HSPVdb. Peptides are either in frame (*y*) or in an ARF (*n*). The position of a SNP is indicated in the column SNP. In addition, the heterozygosity and NetMHC score is given.

CONCLUSIONS

We have shown that selective reduction of the database size by removal of both the non-polymorphic peptide stretches and the non-polymorphic "SNPs" leads to a comprehensive high quality database file dedicated to improving the elucidation of MiHA.

Improvements in the quality and quantity of dbSNP entries, amongst others by the 1000 genomes project (http://www.1000genomes.org), if well controlled, will greatly enhance the use of our database by reporting useful frequencies and removal of spurious frequencies in the current dbSNP releases.

The website (http://srs.bioinformatics.nl/hspv/) provides easy access to relevant information about SNPs by its good annotation and hyperlinks incorporated in the HSPVdb.

ACKNOWLEDGEMENTS

The authors would like to thank David Kloet for initial work on the project. Peter de Koning and Antoinette Teixeira are thanked for peptide synthesis. H.N. was supported by the BioAssist program of the Netherlands Bioinformatics Centre. This research was made possible by the financial assistance of the Landsteiner Foundation for Blood Transfusion Research. The authors declare that they have no conflict of interest.

REFERENCES

Bleakley M, Riddell SR. 2004. Molecules and mechanisms of the graft-versus leukaemia effect. *Nat Rev Cancer* **4**:371.

Edwards NJ. 2007. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol Syst Biol* **3**:102.

Eisenlohr LC, Huang L, Golovina TN. 2007. Rethinking peptide supply to MHC class I molecules. *Nat Rev Immunol* **7**:403–410.

Eng JK, McCormack AL Yates JR 3rd. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**:976-989.

Engelhard VH. 2007. The contributions of mass spectrometry to understanding of immune recognition by T lymphocytes. *Int J Mass Spectrom* **259**:32-39.

Etzold T, Ulyanov A, Argos P. 1996. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* **266**:114-128.

Falkenburg JH, van de Corput L, Marijt EW, Willemze R. 2003. Minor histocompatibility antigens in human stem cell transplantation. *Exp Hematol* **31**:743-751.

Hambach L, Goulmy E. 2005. Immunotherapy of cancer through targeting of minor histocompatibility antigens. *Curr Opin Immunol* **17**:202-210.

Hiemstra HS, Duinkerken G, Benckhuijsen WE, Amons R, de Vries RR, Roep BO, Drijfhout JW. 1997. The identification of CD4 T cell epitopes with dedicated synthetic peptide libraries. *Proc Natl Acad Sci U S A* **94**:10313-10318

Hillen N, Stevanovic S. 2006. Contribution of mass spectrometry-based proteomics to immunology. *Expert Rev Proteomics* **3**:653-664.

Ho O, William R, Green WR. 2006. Alternative Translational Products and Cryptic T Cell Epitopes: Expecting the Unexpected. *J Immunol* **177**:8283-8289.

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. 2004. The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **4**:1985-1988.

Kessler JH, Melief CJ. 2007. Identification of T-cell epitopes for cancer immunotherapy. *Leukemia* **21**:1859-1874.

Marijt WA, Heemskerk MH, Kloosterboer FM, Goulmy E, Kester MG, van der Hoorn MA, van Luxemburg-Heys SA, Hoogeboom M, Mutis T, Drijfhout JW, van Rood JJ, Willemze R, Falkenburg JH. 2003. Hematopoiesis-restricted minor histocompatibility antigens HA-1- or HA-2-specificT cells can induce complete remissions of relapsed leukemia. *Proc Natl Acad Sci U S A* **100**:2742-2747.

Meiring HD, van der Heeft E, ten Hove GJ, de Jong APJM. 2002. Nanoscale LC-MS(n): technical design and applications to peptide and protein analysis. *J Sep Science* **25**:557-568.

Nesvizhskii AI, Vitek O, Aebersold R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **4**:787-797.

Oliveira CC, van Veelen PA, Querido B, de Ru A, Sluijter M, Laban S, Drijfhout JW, van der Burg SH, Offringa R, van Hall T. 2010. The nonpolymorphic MHC Qa-1b mediates CD8+ T cell surveillance of antigen-processing defects. *J Exp Med* **207**(1):207-21.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**:3551-3567.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**:D61-D65.

Reisinger F, Martens L. 2009. Database on Demand - an online tool for the custom generation of FASTA-formatted sequence databases. *Proteomics* **9**(18):4421-4.

Salimi N, Fleri W, Peters B, Sette A. 2010. Design and utilization of epitope-based databases and predictive tools. *Immunogenetics* **62**(4):185-96.

Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, Andersen JS, Mann M. 2007. A mass spectrometry friendly database for cSNP identification. *Nat Methods* **4**:465-466.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**:308-311.

Spierings E, Hendriks M, Absi L, Canossi A, Chhaya S, Crowley J, Dolstra H,Eliaou JF, Ellis T, Enczmann J, Fasano ME, Gervais T, Gorodezky C, Kircher B, Laurin D, Leffell MS, Loiseau P, Malkki M, Markiewicz M, Martinetti M, Maruya E, Mehra N, Oguz F, Oudshoorn M, Pereira N, Rani R, Sergeant R, Thomson J, Tran TH, Turpeinen H, Yang KL, Zunec R, Carrington M, de Knijff P, Goulmy E. 2007. Phenotype frequencies of autosomal minor histocompatibility antigens display significant differences among populations. *PLoS Genet* **3**:e103.

Stepniak D, Wiesner M, de Ru AH, Moustakas AK, Drijfhout JW, Papadopoulos GK, van Veelen PA, Koning F. 2008. Large-scale characterization of natural ligands explains the unique gluten-binding properties of HLA-DQ2. *J Immunol* **180**:3268-3278.

Storb R. 2003. Allogeneic hematopoietic stem cell transplantation--yesterday, today, and tomorrow. *Exp Hematol* **31**:1-10.

The UniProt Consortium. 2008. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **36**:D190-D195.

van Bergen CA, Kester MG, Jedema I, Heemskerk MH, van Luxemburg-Heijs SA, Kloosterboer FM, MarijtWA, de Ru AH, Schaafsma MR, Willemze R, van Veelen PA, Falkenburg JH. 2007. Multiple myeloma-reactive T cells recognize an activation-induced minor histocompatibility antigen encoded by the ATP-dependent interferon-responsive (ADIR) gene. *Blood* **109**:4089-4096.

Chapter 5a

Promoter propagation in prokaryotes

Nucleic Acids Research, 2012, **40**(20): 10032-10040

Mariana Matus-Garcia & Harm Nijveen and Mark W. J. van Passel

Supplemental material is available at:
http://nar.oxfordjournals.org/content/40/20/10032/suppl/DC1

Mariana Matus-Garcia & Harm Nijveen share first authorship

ABSTRACT

Transcriptional activation or 'rewiring' of silent genes is an important, yet poorly understood, phenomenon in prokaryotic genomes. Anecdotal evidence coming from experimental evolution studies in bacterial systems has shown the promptness of adaptation upon appropriate selective pressure. In many cases, a partial or complete promoter is mobilized to silent genes from elsewhere in the genome. We term hereafter such recruited regulatory sequences as Putative Mobile Promoters (PMPs) and we hypothesize they have a large impact on rapid adaptation of novel or cryptic functions. Querying all publicly available prokaryotic genomes (1,362) uncovered over 4,000 families of highly conserved PMPs (50 to 100 nt long with ≥ 80% nt identity) in 1,043 genomes from 424 different genera. The genomes with the largest number of PMP families are *Anabaena variabilis* (28 families), *Geobacter uraniireducens* (27 families), and *Cyanothece* PCC7424 (25 families). Family size varied from 2 to 93 homologous promoters (in *Desulfurivibrio alkaliphilus*). Some PMPs are present in particular species but some are conserved across distant genera. The identified PMPs represent a conservative dataset of very recent or conserved events of mobilization of non-coding DNA and thus they constitute evidence of an extensive reservoir of recyclable regulatory sequences for rapid transcriptional rewiring.

INTRODUCTION

Transcriptional rewiring is a term used for defining the modification of transcriptional circuits over evolutionary time, due to changes in transcription factors (TFs) and/or cis-regulatory elements. This concept has been widely used in studies of eukaryotic transcription circuits (Tuch et al. 2008), but much less in prokaryotic systems, mainly because the extent of the phenomenon in bacteria is presently unknown (Perez and Groisman 2009; Wang et al. 2011).

However, transcriptional rewiring may actually play an important role in prokaryotic genome evolution given the large turnover of gene functions. Indeed the prevalence of gene acquisition through horizontal gene transfer (HGT) (Ochman et al. 2000; Popa et al. 2011; Treangen and Rocha 2011) and gene loss from deletion events (van Passel et al. 2007; van Passel et al. 2008) generates highly dynamic genomes that differ even between closely related species or strains. As examples of such a large turnover of genes, it has been estimated that 61 genomes of *Escherichia coli* strains share only about 20% of gene functions (Lukjancenko et al. 2010).

Transcriptional rewiring can result in activation of silent genes, such as HGT-derived genes without a compatible promoter (Pal et al. 2005), or in modification of the expression of already present genes. Such activation requires as a first step the evolution of a functional promoter, *i.e.* -10, -35 boxes and TF binding sites that can be recognized by the cell's transcriptional machinery (Browning and Busby 2004). In principle, a promoter could evolve by two different mechanisms. It can evolve *de novo* by the creation of cis-regulatory elements through point

mutations and indels (Lercher and Pal 2008). Alternatively, it can evolve in a single 'quantum leap' through the recruitment or mobilization of already existing promoters from elsewhere in the genome (Stavrinides et al. 2006).
Experimental evolution studies in *Pseudomonas putida* (Kasak et al. 1997), *Lactoccocus lactis* (Bongers et al. 2003; de Visser et al. 2004) and *Escherichia coli* (Posfai et al. 2006; Lee and Palsson 2010; Stoebel and Dorman 2010) have found promoter recruitment to be the main mechanism driving transcriptional activation or rewiring of silent genes, through mobilization of partial or complete promoters by transposable elements (Zhang and Saier 2009).
Furthermore, recent advances in understanding the function of DNA repeats in intergenic regions have shown that they can have important regulatory roles in transcription or translation (Delihas 2011); and given their ability to propagate, DNA repeats can also be involved in transcriptional rewiring. Miniature inverted terminal repeat elements (MITEs) are nonautonomous mobile elements, that is, they only transpose if a suitable transposase is provided in *trans* by an autonomous IS element. Examples of MITEs that can influence transcription are the *Neisseria* CREE element (Snyder et al. 2009; Siddique et al. 2011) and the *Yersinia* ERICS (De Gregorio et al. 2005), both of which carry partial promoters at their termini.
Based on these observations, it seems that intragenomic promoter propagation could represent a major force driving transcriptional activation or rewiring in prokaryotes. In the present study, the extent of promoter propagation in archaea- and bacteria was assessed by *in silico* analysis of all publicly available genomes. Evidence for promoter propagation events was found in >4,000 families of conserved homologous sequences upstream of non-homologous CDSs. These 'Putative Mobile Promoters' (PMPs) present examples of reported insertion sequences and riboswitches, but notably also a large fraction of novel families of dynamic elements with potential influence on transcription. We hypothesise that PMPs may represent a vast recyclable reservoir of regulatory potential for rapid transcriptional recruitment or rearrangement.

MATERIAL AND METHODS

Identification of intra-genomic promoter propagation

To identify putative mobile promoters in a bacterial genome we looked for conserved homologous sequences upstream of non-homologous CDSs (Figure 1). The promoter of each CDS was assumed to be contained in the first 150 to 100-nt upstream of the translation start site (TLS) of predicted transcriptional units. This assumption builds on the finding that bacterial promoters are relatively compact with 100-nt regions generally containing the regulatory signals needed for initiating transcription (Perez and Groisman 2009). Furthermore, those regulatory signals are usually located immediately upstream of CDSs. For example, the majority of transcriptional start sites in *Escherichia coli* K12 are located between 20 to 40 nucleotides from the TLS, and most of the TF-binding

sites are located 50-nt upstream of the transcriptional start site (Mendoza-Vargas et al. 2009). Therefore it can be reasonably assumed that the method deals with sequences probably involved in transcriptional regulation.

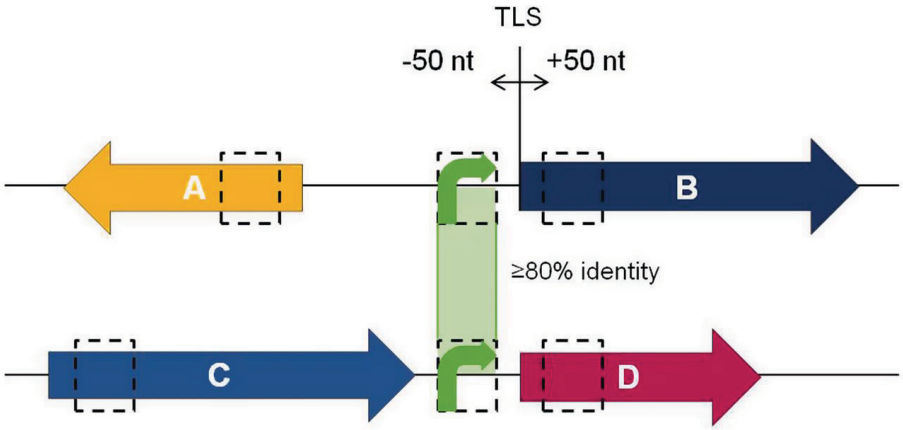


Figure 1. Identification of PMPs. Dashed boxes represent 100-nt defined promoters (green arrows), downstream CDSs (dark blue and pink arrows) and upstream CDSs (orange and blue arrows) used for BLAST alignments. Those regions were taken ±50 nt of the translation start site (TLS) of the downstream CDSs. Two promoters are considered mobile if they align over >50 nt with at least 80% identity (green shadow), while their upstream CDSs (A and C) and downstream CDSs (B and D) do not align.

We took 100-nt fragments from all promoters and CDSs found in a genome starting at 50-nt upstream or downstream respectively of the TLS as depicted (Figure 1). The sequences were extracted with an in-house developed Perl script using the annotation (.ptt) and the FASTA files (.fna) of 1362 complete prokaryote genomes (archaea- and eubacteria; 971 species; 503 genera; see Supp. Table 1 for complete list) reported at the NCBI website (May 2011). The collected sequences from different chromosomes and/or plasmids of the same genome were stored in one file and formatted as a BLAST database. The BLAST (Altschul et al. 1997) alignments were performed within each genome using an E-value cut-off of 0.0001 and the filter for low complexity regions off. A hit between promoters was considered relevant if the alignment was at least 50-nt long with 80% identity (i.e., at least 40 out of 50 nucleotides were identical) while all hits between coding regions were considered indicative of homology. All filtered pair-wise hits were clustered with the netclust (score cut-off of zero) program (Kuzniar et al. 2010) to obtain the unfiltered families (we call pre-clusters) of homologous sequences per genome. A pre-clusters was discarded if (i) it contained both promoters and CDSs, since these sequences could represent misannotated TLSs, or (ii) the whole gene was duplicated (promoter region and CDS), since we are interested only in promoter mobilization. Pre-clusters passing the filters became families of PMPs. In each family, the promoter showing homology to the most members was

selected as representative. If the representative was homologous to all members in its family, then it was said to be a central node and it indicated the presence of a highly conserved core in the family.

Identification of inter-genomic promoter propagation

CD-HIT-EST (Huang et al. 2010) was used to cluster all representatives at 80% identity over 50 nt (program parameters: -c 0.8 –G 0 –aL 50). The clustering removed redundancy in the dataset and identified PMPs in different strains of the same species, different species of the same genus or bacteria from different genera.

Control dataset: randomized sequences

To estimate the number of duplicated promoters that one could expect to find by chance, we generated a mock dataset with shuffled sequences having the same promoter' and CDS' nucleotide compositions for each genome. Sequences were re-shuffled 10 to 20 times with an in-house developed Perl script, and then run through the pipeline.

Functional analyses of the propagated promoters

Quantitative analyses were carried out to investigate the incidence, conservation and possible function of mobile promoters. The non-redundant dataset was used to query RFAM (Gardner et al. 2009), IS Finder (Siguier et al. 2006) and published MITEs datasets (Snyder et al. 2009; Delihas 2011) to assess how many of the identified promoters are actually known RNA regulatory elements, insertion sequences or non-autonomous mobile elements. The cmsearch program of the INFERNAL suite (Nawrocki et al. 2009) was used to search against the 1,973 RFAM calibrated models (14 June 2011 release) with the trusted cut-off (--tc). The IS Finder web server was used to search for reported IS elements with an *E*-value cut-off of 0.0001 and with filter for low complexity regions off. To find the more distant members of each PMP family and thus gain insight into the propagation dynamics of PMPs, we extended the families with all BLAST hits having an *E*-value < 0.0001 that did not pass the alignment length and identity filters.

Finally a comparison of PMPs present in *E. coli* strains was performed to check for inter-strain variability.

Pipeline

A pipeline script was programmed in Perl to automate every step of the analysis, except for the use of IS Finder. The pipeline runs in a Linux environment and it requires the data and supporting programs to be installed locally. Please contact the authors for the suite of scripts and instructions.

RESULTS

Identification of intra-genomic promoter propagation

Putative mobile promoters were identified as highly similar stretches of non-coding DNA located in promoter regions of non-homologous genes in a species (Figure 1). All promoter sequences with minimal length of 150 nt upstream of the start codon (1,142,064) were mined from 1,362 prokaryotic genomes and formed 11,821 pre-clusters. Over 60% (7,366) of them also shared homology in their corresponding downstream CDSs, and thus cannot be considered as only promoter duplications. This strong reduction to 4,455 families indicates that most of the highly conserved duplicated promoters in these bacterial genomes are in fact part of complete gene duplications. We also filtered out cases of homology in the neighbouring upstream CDS and were left with a final dataset of 4,071 families (13,111 sequences; see Supp. Data for FASTA sequences of identified PMPs). Among the discarded data we found several cases (47% = 180/381 pre-clusters) in which the conserved promoters were actually long terminal inverted repeats from transposases present in multiple copies in the genome (for example Supp. Figure 1).

Analysis of the family of 10 members in *Treponema brennaborense* DSM-12168 (Figure 2 and Supp. Table 2) showed that the promoters are highly similar to each other (average identity of 95%) over a large stretch (average length of 84 nt). Upon closer inspection it was found that sequence conservation starts around position -5 upstream of the TLS and extends up to position -120 with less conserved sequences up to -170 nt.

Identification of inter-genomic promoter propagation

Redundancy in the dataset caused by over-representation of certain bacterial clades in the genomes database (e.g. *E. coli*) was not purged from the beginning because it was of interest to identify recent promoter propagation events across strains of the same species. To estimate the level of redundancy in the results and to pinpoint cases of putative mobile promoters across different species or genera, all identified duplicated promoters were clustered together (see Supp. Table 3 and Supp. Data for FASTA sequences of non-redundant inter-genomic PMPs). From the 4,074 families in the final dataset, 3,216 non-redundant families were formed of which 87% (2,791/3,216) were formed by single representatives. The rest consisted of homologous promoters between different strains of the same species (168 families), different species of the same genus (146 families) or different genera (75 families). The latter are of particular interest since they could represent cases of HGT-derived promoters present in distant species. For example a putative mobile promoter was found upstream of eight different CDSs in *Herpetosiphon aurantiacus* ATCC 23779, *Carboxydothermus hydrogenoformans* Z 2901, *Deinococcus maricopensis* DSM 21211, and *Thermotoga lettingae* TMO (Figure 3 and Supp. Table 4). There were also 36 families formed by representatives from

the same genome that were not grouped together in the pipeline (due to identity or length filters) but are indeed homologous.

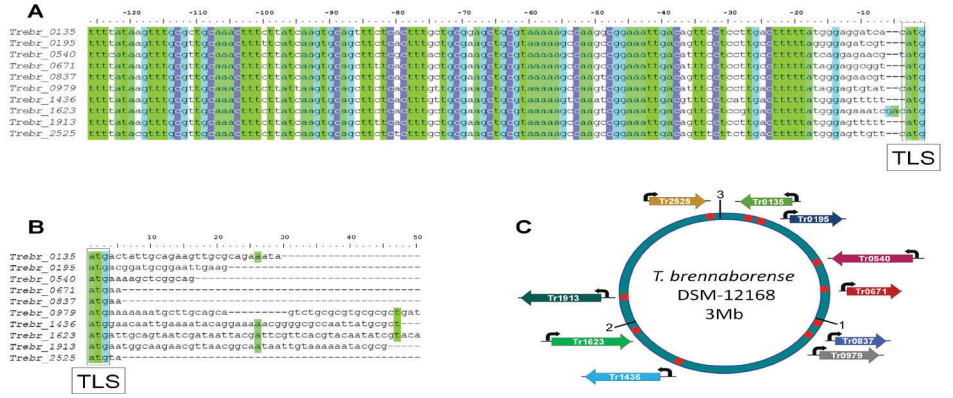


Figure 2. Putative Mobile Promoter regulating 10 non-homologous CDSs. (**A**) Alignment of the multiple copies of a PMP in *T. brennaborense*. A highly conserved core can be observed in the region -120 to -5 upstream of the TLS of the downstream CDSs. Colour blocks represent conserved residues. The location of the TLS is indicated. (**B**) Alignment of the CDSs downstream of the PMP. No sequence conservation is observed. (**C**) Location of the PMPs (black arrows) along the 3 Mbp circular chromosome (Mbp are marked as 1, 2 and 3 in the figure). The orientation of each gene is depicted according to the genome annotation.

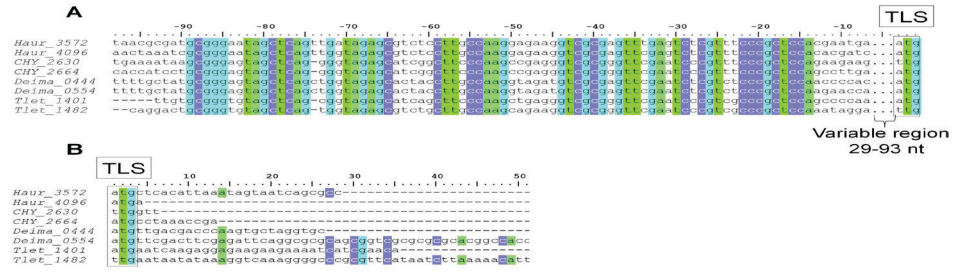


Figure 3. Alignment of a Putative Mobile Promoter present in four distant species. (**A**) Alignment of the promoter region of eight CDSs in *Herpetosiphon aurantiacus* ATCC23779, *Carboxydothermus hydrogenoformans* Z 2901, *Thermotoga lettingae* TMO and *Deinococcus maricopensis* DSM2121. Colour blocks show nucleotide sequence conservation. A variable region of non-conserved sequence (29-93 nt long) was observed in all cases between the TLS and the PMP. (**B**) Alignment of the coding sequence of the eight downstream CDSs. No sequence conservation is observed.

Control dataset: randomized sequences

Using randomized sequences with the same nucleotide composition (mock data), 1,457 pre-clusters were obtained through BLAST and netclust (149,107 in real data). All of the pre-clusters were formed by a mixture of CDSs and promoters

sequences, since both types of sequences are used for the clustering in the pipeline. Therefore no pre-cluster made it through to the final families dataset (11,821 did in the real data). Looking at the genomes that were particularly enriched with random pre-clusters, we found four genomes with over 20 (Supp. Table 5). All such genomes have a skewed base composition (<30 or >70 %GC), which could explain the high number of random sequence conservation. This is supported by a plot of %GC versus number of families (Supp. Figure 2). In the real dataset, none of these genomes had a particularly high count of families (all ≤10 families) and the number of families was not correlated to the GC content of the DNA molecule (Supp. Figure 2). The top five genomes with the largest number of families in the dataset (all ≥21 families) had zero or one family in the mock data.

Quantitative analysis of the PMPs distribution

4,074 families were mined from 1,043 prokaryotic genomes representing 424 genera. The genera with most families were the ones with more sequenced representatives, e.g. *Clostridium* (149 families; 31 genomes), *Escherichia* (141 families; 31 genomes), *Streptococcus* (125 families; 52 genomes), and *Bacillus* (104 families; 37 genomes). Normalizing the number of families by the number of genomes in the database showed that four cyanobacterial genera and one bacteroidetes had the highest enrichment of families relative to the number of available genomes. The species with the largest number of PMP families are *Anabaena variabilis* ATCC29413 (28 families), *Geobacter uraniireducens* Rf4 (27 families), *Cyanothece* PCC7424 (25 families), *Trichodesmium erythraeum* IMS101 (22 families), and *Psychromonas ingrahamii* 37 (21 families). All five species have large circular chromosomes (Supp. Table 5) suggesting that genome size could be correlated with the number of duplicated promoters, a similar correlation has been reported for gene paralogs (Gevers et al. 2004) and regulatory potential (Konstantinidis and Tiedje 2004). However no correlation was observed when plotting the size of the DNA molecule (chromosome or plasmid) versus the number of families, nor the number of mock pre-clusters (Supp. Figure 3). The same result was observed when plotting the total genome size versus the number of families (data not shown).

About 80% of the analysed genomes contain less than 6 families of duplicated promoters (78% = 812/1043 genomes; see Figure 4A) and the majority have only one family. This overall low count of propagated promoters suggests that either mobilized promoters diverge very fast and the present methodology is too conservative to find more cases, or that promoter mobilization independent of CDS duplication is a rare event.

Small family sizes were obtained with the majority having only two members (68% = 2,771/4,074 families; see Fig 4B). These pairs were on average highly conserved (mean identity of 92%, see Fig 5A) and the majority were of the minimal allowed alignment length (50 nt, see Fig 5B). Interestingly, the most frequent case was that of identical promoters, which again implies the pipeline is

finding predominantly very recent or conserved duplications.
The largest family (93 promoters) was found in the anaerobic sulphur-reducer *Desulfurivibrio alkaliphilus* AHT2.

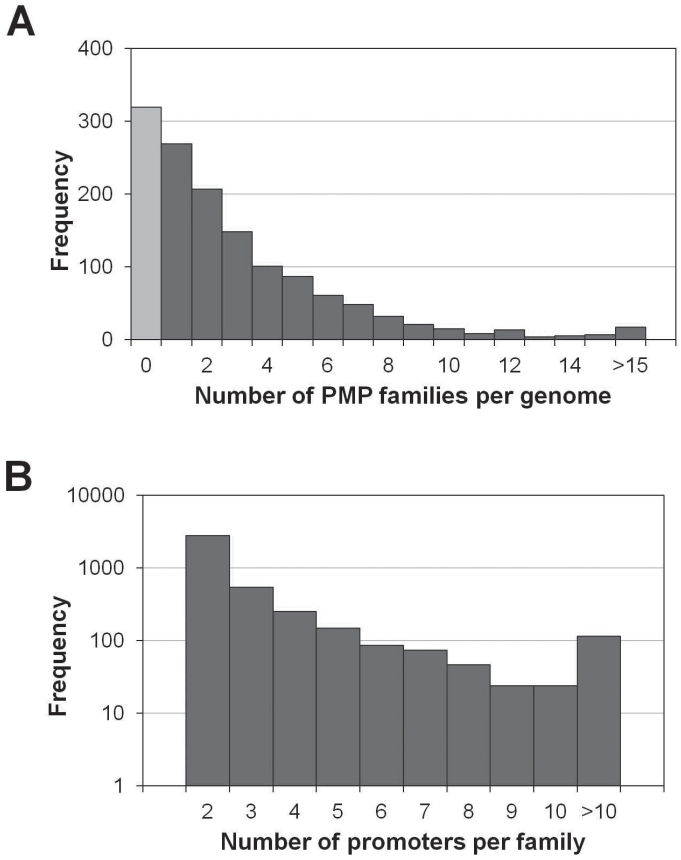


Figure 4. Quantification of Putative Mobile Promoter propagation in prokaryotic genomes. (**A**) Number of families per genome (total= 4,074 families in 1,043 genomes). (**B**) Number of promoters per family (total= 13,111 promoters in 4,074 families); please note the logarithmic scale on the y-axis.

Search of riboswitches, IS elements and MITES

Riboswitches and insertion sequences (IS) are known elements with possible regulatory functions. In order to examine the fraction of PMPs that are in fact such reported elements; we queried representative sequences from the non-redundant data set (3,216 sequences) against the RFAM and IS Finder databases.
Searching the RFAM database resulted in 125 hits (~4% =125/3,216 representatives) with 33 RNA models of RFAM (out of 1,973 present in the database). The most frequent hit was with tRNAs (42/125 hits), which are known integration sites for genomic islands (Ou et al. 2006).

The method effectively purged IS elements from the dataset by restricting sequence conservation only in the promoter regions and not in their neighbouring CDSs. However IS elements can leave behind direct repeats when they excise and insert in another location. Searching against the IS Finder web server to find traces of similarity to IS elements, 210 hits (~7% =210/3,216) with 177 different IS were retrieved. *Methylobacterium extorquens* AM1 had most hits with the database (5/14 families).

Two PMPs had hits both with RNA-regulatory elements and IS elements. One is a pair present in *Stenotrophomonas maltophilia* that presented similarity to the mraW RNA motif, associated with peptidoglycan synthesis (RFAM, http://rfam.sanger.ac.uk), and to the ISStma8 transposase (IS110 family). The other doublet is present in *Glaciecola agarilytica* and was similar to the antisense RNA-OUT that regulates transposition and the ISPat1 (IS4 family). Thus, the resemblance to IS elements could provide the RNA elements with mobility. This is interesting since the mechanism by which riboswitch families expand or shrink is presently unknown. However it can be anticipated that the dynamics of mobile elements (*e.g.* insertion sequences, transposases, etc.) can result in different frequencies of the RNA elements, *e.g. Streptomyces coelicolor's* genome has nine copies of the adenosyl-cobalamin riboswitch (Ado-CBL) while *Streptomyces avermitilis'* has four.

The fact that the dataset had a low count of reported riboswitches and IS elements (together ~11% of families) indicates that our methodology finds mainly new mobile regulatory elements.

To investigate the occurrence of MITEs in the dataset, all representatives were searched against a database of 5' UTR CREE elements (Snyder et al. 2009). None was found in the dataset. Manual checking confirmed that such repeats were excluded early in the pipeline because they are present both in promoters and CDSs regions.

Functional categories of CDSs downstream of PMPs

To analyse if the putative mobile promoters that we find are biased towards certain functional classes of genes, the Cluster of Orthologous Groups (COG) (Tatusov et al. 1997) classification from all downstream CDSs was obtained. With respect to the encoded product, most of the genes encode hypothetical proteins (4,809/13,111 CDSs) followed by transposases (295/13,111 CDSs) and GCN5-related N-acetyltransferase (57/13,111 CDSs). Only in a minimal fraction of the families (3% = 130/4,074 families) all members of the same family belong to the same COG. These could represent genes involved in the same metabolic pathways that would benefit from co-regulation.

These data together imply that little information is available for the CDSs found in our study, which is in accordance with our hypothesis that putative mobile promoters could be involved in recent events of transcriptional rewiring of species-specific genes rather than housekeeping functions.

Strain	No. of PMP families	Total number of sequences
E. coli 536	4	12
E. coli 55989	4	12
E. coli APEC O1	4	8
E. coli ATCC 8739	5	27
E. coli B REL606	6	26
E. coli BL21 Gold DE3 pLysS AG	5	21
E. coli BW2952	6	25
E. coli CFT073	3	7
E. coli E24377A	3	12
E. coli ED1a	6	15
E. coli HS	5	13
E. coli IAI1	3	9
E. coli IAI39	2	7
E. coli K 12 substr DH10B	3	9
E. coli K 12 substr MG1655	5	24
E. coli K 12 substr W3110	5	25
E. coli O103 H2 12009	6	22
E. coli O111 H 11128	4	18
E. coli O127 H6 E2348 69	6	13
E. coli O157 H7 EC4115	7	16
E. coli O157 H7 EDL933	6	16
E. coli O157 H7 Sakai	3	8
E. coli O157 H7 TW14359	7	16
E. coli O26 H11 11368	5	19
E. coli O55 H7 CB9615	3	8
E. coli S88	7	15
E. coli SE11	8	19
E. coli SMS 3 5	2	4
E. coli UMN026	3	9
E. coli UTI89	3	6

Table 1. Differences in PMP family number and size in 30 strains of *Escherichia coli*.

Case study: Escherichia coli

Rapid propagation of the putative mobile promoters throughout genomes could result in different frequencies of these promoters in closely related strains. An example of intra-species variation was analysed in *Escherichia coli*, which is represented in the database by 30 sequenced strains. It was found that even between closely related strains there were substantial differences in the number of families and/or number of promoters in the families (Table 1). Families were found in all 30 reported genomes however the numbers varied from 2 to 8. Differences were found even between isolates of the same strain, for example in *E. coli* K12 MG1655 (five families) and *E. coli* K12 DH10B (three families). To validate that the different counts are not an artefact of the set identity and length thresholds, promoter families were made again but taking into account all BLAST hits. Differences in abundance of families and number of promoters were found again thus showing that the PMPs do have different frequencies in closely related strains. For example Table 2 shows the distribution of a PMP across different *Enterobacteriales (E. coli, Salmonella enterica, Shigella boydii and Yersinia pestis)*. The downstream CDSs of the PMP are classified into a large variety of COGs and the degree of sequence conservation is also variable. Diverged copies of the PMP are indicated with gray cells in the table, and conserved copies with brown cells. All *E. coli* CDSs downstream of PMPs were checked to determine the abundance of horizontal gene transfer, by using a dataset of identified HGT events (Popa et al. 2011). It was found that about 25% of the CDSs in our dataset present evidence of HGT (Chi square test at p-value=0.0001), which is about the same as for all *E. coli* genes (30%). Therefore our dataset of PMPs is involved both in transcriptional activation events for HGT-genes but primarily in transcriptional rewiring of already existing functions. Another interesting observation is that in some cases the number of families and family members did not change or only very little, *e.g. E. coli* O157 family (Table 2), while in other cases the total number of promoters increased dramatically, *e.g.* the family in *Y. pestis* grew from 13 to 100 promoter members (see Supp. Table 6 for complete list of PMP families, members, riboswitches and IS elements per genome). Such difference in occurrence and conservation could provide information on the mechanism by which the promoters are being mobilized. A promoter with tens or hundreds of copies in a genome could well represent a non-autonomous mobile element that is copied by an active transposase, while a promoter present in two or three copies could be result of random duplication through homologous or non-homologous recombination.

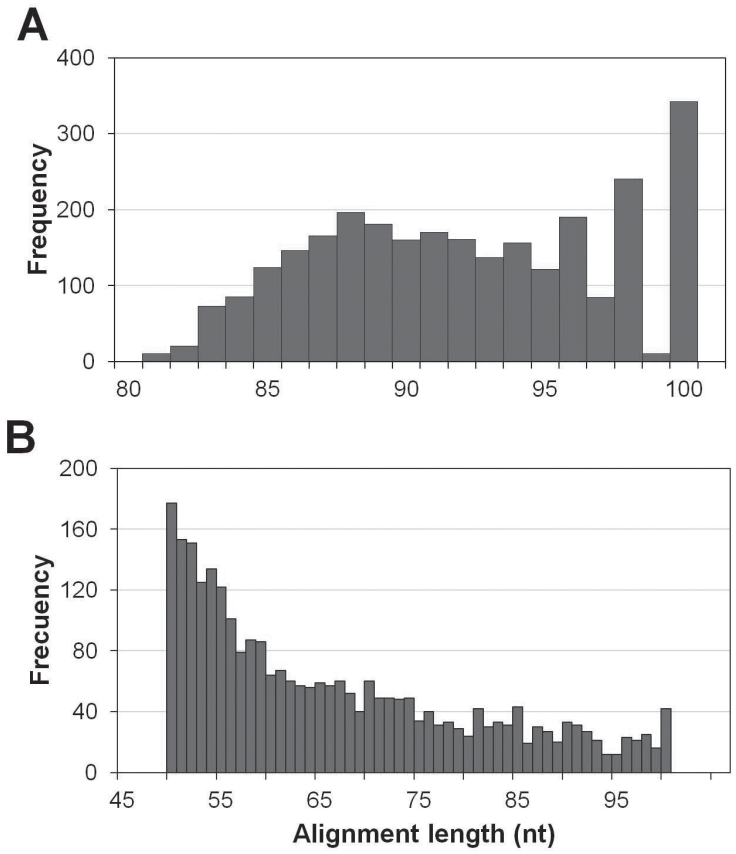


Figure 5. Two-members families features (2,771/4,074 families). (**A**) Identity distribution (mean=92%). (**B**) Alignment length distribution (mean=66 nt).

DISCUSSION

Treangen *et al.* (2009) provide an operational definition of DNA repeats based on three properties of the copies: i) the distance between them, ii) the similarity level and iii) the length over which they align. Analyses of such properties have produced the guideline that exact repeats larger than 25 nt are statistically significant in most prokaryotic genomes (Treangen et al. 2009). Since the present study required alignments of at least 50 nt with 80% identity, the dataset presented is a conservative investigation of the repeats found in promoter regions throughout the bacterial and archaeal domains. We showed that neither the length nor composition of the DNA molecules is correlated to the presence of putative mobile promoters. Our analysis pipeline did not find any family of mobile promoters in a control randomized dataset (Supp. Table 5). Therefore we are confident that the data presented in this study indeed represents statistically significant events of promoter propagation.

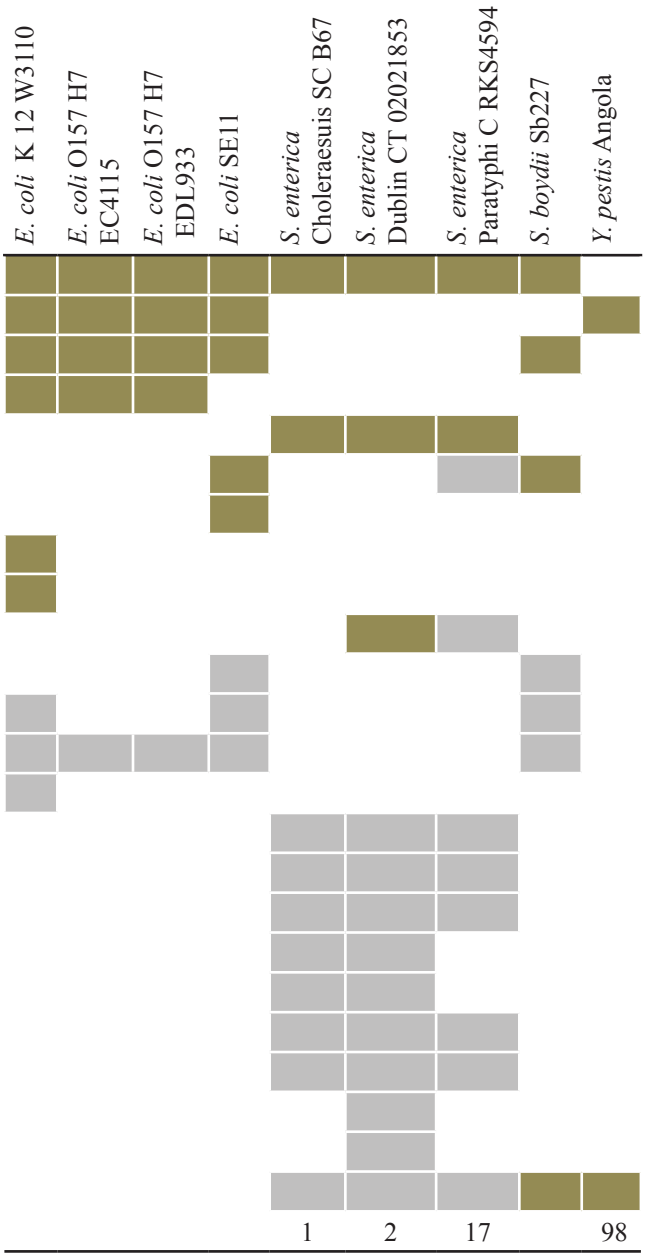
Bacteria seem to employ various mechanisms to be able to re-use promoter sequences instead of having to evolve them *de novo*. Based on reported literature and inspection of the dataset, we propose that promoters can be mobilized through four main mechanisms: i) Mobile elements, either as part of the terminal inverted repeats (Bongers et al. 2003), or linked to them (Stavrinides et al. 2006); ii) Non-autonomous mobile elements, emulating terminal inverted repeats (Delihas 2011); iii) Random duplications mediated by recombination processes; iv) Horizontal gene transfer, which actually is the result of mobile elements (*e.g.* conjugative plasmids) or duplications (e.g. minimal mobile elements). Families of promoters that were conserved along with an upstream CDS are probably examples of mobile elements (transposases) that carry promoters (see Supp. Figure 1) in their termini. Families that grew dramatically when all BLAST hits were taken into account probably represent groups of non-autonomous mobile elements or scars from autonomous mobile ones. Pairs found in single species are probably examples of random promoter duplications resulting from homologous or non-homologous recombination. Families present in bacteria from different species or genera could represent HGT-derived promoters (Figure 3 and Supp. Table 4), potentially capable of being functional in a broad host range. Although there are no reported cases of HGT-derived promoters, it is a plausible scenario since any type of DNA can undergo lateral transfer (Ragan and Beiko 2009).

This rapid integration of novel gene functions probably is an important factor in the success of horizontal gene transfer and the rapid adaptation to novel niches. It is presently unknown to which extent HGT-derived genes come with a promoter that can be used straightaway. However there are indications that such a promoter-CDS co-transfer is unlikely to occur since expression of the novel gene can be deleterious to fitness or even lethal if the novel CDS product is toxic or poses gene dosage problems (Sorek et al. 2007), plus there is an inherent limitation to HGT regarding the length of simultaneously transferred DNA. Therefore, recycling of appropriate promoters for HGT-derived CDSs seems to be a plausible, economic and biologically significant event in the integration of novel gene functions. This is in agreement with the finding that the evolutionary rate of non-coding upstream sequences is higher for the most recent HGT-derived CDSs in *E. coli* K12 (Lercher and Pal 2008).

The results also show how bacteria could recycle genetic material not only at the CDS level for generating paralogs in the process of neofunctionalization, but also in non-coding regions to generate (novel) families of regulatory sequences. Since mainly small PMP families are identified, it seems that either they diverge very fast or family expansion is uncommon. Family expansion to include all BLAST hits of the PMP provided examples of both cases. While the doublets (families of two promoters) were highly conserved (~92% identity, Figure 5), the larger families already presented many variations near the TLS (see Figures 2 and 3 for examples). This could be an illustration of how a generic mobile promoter adapts to produce different transcriptional responses in the downstream CDSs,

		E. coli 55989	*E. coli* BW2952	*E. coli* HS	*E. coli* IAI1
COG0260E	PepB. Aminopeptidase B.	■	■		■
COG1048C	YbhJ. Predicted hydratase.	■	■	■	
COG0191G	FbaA. Fructose-bisphosphate aldolase.	■	■	■	■
COG1690S	YkfJ. Hypothetical protein.				
COG0300R	Short-chain dehydrogenases.				
COG2141C	Hypothetical protein.	■			■
COG0493ER	Putative oxidoreductase.	■			■
COG0667C	YajO. 2-carboxybenzaldehyde reductase.		■	■	
COG5569S	CusF. Periplasmic copper/silver binding protein.		■	■	
COG2116P	Formate/nitrite family of transporters.				
COG1966T	CstA. Putative carbon starvation protein.			▒	■
COG1249C	Lpd. Lipoamide dehydrogenase.	▒	▒	▒	▒
COG0277C	YdiJ. Oxidoreductase FAD-binding protein.	▒	▒	▒	▒
COG0286V	DNA methyltransferase M.		▒		
COG0121R	Predicted glutamine amidotransferase.				
COG1349KG	Transcriptional regulators of sugar metabolism.				
COG0567C	2-oxoglutarate dehydrogenase complex.				
COG0813F	Purine nucleoside phosphorylase.				
COG0246G	Mannitol-1-phosphate/altronate dehydrogenases.				
COG0369P	Inorganic ion transport and metabolism.				
COG0166G	Glucose-6-phosphate isomerase.				
COG2844O	UTP:GlnB (protein PII) uridylyltransferase.				
COG0129EG	Dihydroxyacid/phosphogluconate dehydratase.				
N.A.	-				
Other COG				1	

Table 2. Occurrence of a PMP family in different strains of *E. coli, S. enterica, S. boydii* and *Y. pestis*. The different strains are shown in the columns while rows stand for COG annotations. Brown cells show highly conserved copies of the PMP and gray cells correspond to diverged copies (below length and identity thresholds). Differences in PMP frequencies and conservation can be observed between different genera, species and strains.


E. coli K 12 W3110
E. coli O157 H7 EC4115
E. coli O157 H7 EDL933
E. coli SE11
S. enterica Choleraesuis SC B67
S. enterica Dublin CT 02021853
S. enterica Paratyphi C RKS4594
S. boydii Sb227
Y. pestis Angola
1
2
17
98

providing thus flexibility in the type of regulation it provides. This also indicates that most probably the doublets represent the most recent cases of promoter propagation, which is supported by the fact that identical promoters are the most common case (Figure 5). Finally, it can also be argued that the fast divergence of PMPs families also prevents genomic instability by quickly reducing the chance of recombination between identical copies. This could explain why we find highly conserved PMP families at a low frequency in all analysed genomes (on average 3 families per genome) with the conservative methodology we followed. It will be interesting to determine which proportion of the PMPs are transcriptional activators, down-regulators or even silencers, and if their function lies at the transcriptional or post-transcriptional level.

ACKNOWLEDGEMENTS

This work is dedicated to the memory of Professor Jack A.M. Leunissen, one of the first Dutch bioinformaticians.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Bongers RS, Hoefnagel MH, Starrenburg MJ, Siemerink MA, Arends JG, Hugenholtz J, Kleerebezem M. 2003. IS981-mediated adaptive evolution recovers lactate production by ldhB transcription activation in a lactate dehydrogenase-deficient strain of Lactococcus lactis. *J Bacteriol* **185**(15): 4499-4507.

Browning DF, Busby SJ. 2004. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**(1): 57-65.

De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP. 2005. Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J Bacteriol* **187**(23): 7945-7954.

de Visser JA, Akkermans AD, Hoekstra RF, de Vos WM. 2004. Insertion-sequence-mediated mutations isolated during adaptation to growth and starvation in Lactococcus lactis. *Genetics* **168**(3): 1145-1157.

Delihas N. 2011. Impact of small repeat sequences on bacterial genome evolution. *Genome Biol Evol* **3**: 959-973.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**(Database issue): D136-140.

Gevers D, Vandepoele K, Simillon C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* **12**(4): 148-154.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**(5): 680-682.

Kasak L, Horak R, Kivisaar M. 1997. Promoter-creating mutations in Pseudomonas putida: a model system for the study of mutation in starving bacteria. *Proc Natl Acad Sci U S A* **94**(7): 3134-3139.

Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* **101**(9): 3160-3165.

Kuzniar A, Dhir S, Nijveen H, Pongor S, Leunissen JA. 2010. Multi-netclust: an efficient tool for finding connected clusters in multi-parametric networks. *Bioinformatics* **26**(19): 2482-2483.

Lee DH, Palsson BO. 2010. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microbiol* **76**(13): 4158-4168.

Lercher MJ, Pal C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* **25**(3): 559-567.

Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* **60**(4): 708-720.

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B et al. 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS ONE* **4**(10): e7526.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**(10): 1335-1337.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**(6784): 299-304.

Ou HY, Chen LL, Lonnen J, Chaudhuri RR, Thani AB, Smith R, Garton NJ, Hinton J, Pallen M, Barer MR et al. 2006. A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res* **34**(1): e3.

Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**(12): 1372-1375.

Perez JC, Groisman EA. 2009. Transcription factor function and promoter architecture govern the evolution of bacterial regulons. *Proc Natl Acad Sci U S A* **106**(11): 4319-4324.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**(4): 599-609.

Posfai G, Plunkett G, 3rd, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M et al. 2006. Emergent properties of reduced-genome Escherichia coli. *Science* **312**(5776): 1044-1046.

Ragan MA, Beiko RG. 2009. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci* **364**(1527): 2241-2251.

Siddique A, Buisine N, Chalmers R. 2011. The transposon-like Correia elements encode numerous strong promoters and provide a potential new mechanism for phase variation in the meningococcus. *PLoS Genet* **7**(1): e1001277.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**(Database issue): D32-36.

Snyder LA, Cole JA, Pallen MJ. 2009. Comparative analysis of two Neisseria gonorrhoeae genome sequences reveals evidence of mobilization of Correia Repeat Enclosed Elements and their role in regulation. *BMC Genomics* **10**: 70.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**(5855): 1449-1452.

Stavrinides J, Ma W, Guttman DS. 2006. Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. *PLoS Pathog* **2**(10): e104.

Stoebel DM, Dorman CJ. 2010. The effect of mobile element IS10 on experimental regulatory evolution in Escherichia coli. *Mol Biol Evol* **27**(9): 2105-2112.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* **278**(5338): 631-637.

Treangen TJ, Abraham AL, Touchon M, Rocha EP. 2009. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol Rev* **33**(3): 539-571.

Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7**(1): e1001284.

Tuch BB, Li H, Johnson AD. 2008. Evolution of eukaryotic transcription circuits. *Science* **319**(5871): 1797-1799.

van Passel MW, Marri PR, Ochman H. 2008. The emergence and fate of horizontally acquired genes in Escherichia coli. *PLoS Comput Biol* **4**(4): e1000059.

van Passel MW, Smillie CS, Ochman H. 2007. Gene decay in archaea. *Archaea* **2**(2): 137-143.

Wang L, Wang FF, Qian W. 2011. Evolutionary rewiring and reprogramming of bacterial transcription regulation. *J Genet Genomics* **38**(7): 279-288.

Zhang Z, Saier MH, Jr. 2009. A novel mechanism of transposon-mediated gene activation. *PLoS Genet* **5**(10): e1000689.

Chapter 5b

Promoter reuse in prokaryotes

Mobile Genetic Elements, 2012, **2**(6): 6-8

Harm Nijveen, Mariana Matus-Garcia, Mark W.J. van Passel

Commentary to Matus-Garcia et al. 2012, chapter 5a of this thesis

ABSTRACT

Anecdotal evidence shows promoters being reused separate from their downstream gene, thus providing a mechanism for the efficient and rapid rewiring of a gene's transcriptional regulation. We have identified over 4,000 groups of highly similar promoters using a conservative sequence similarity search in all fully sequenced prokaryotic genomes. About 6% of those groups are shared between bacteria from different taxonomic depth, including different genera, families, orders, classes and even phyla. Database searches against known mobile elements and RNA motifs have indicated that regulatory motifs such as riboswitches could be moved around on putative mobile promoters.

Introduction

Reuse of protein coding DNA sequences through gene duplication and horizontal gene transfer is a well-known and profound innovative force in nature (van Passel et al. 2008); in sharp contrast to this, the mobility of a gene's transcription regulatory function encapsulated in its promoter region is much less known. There are a few well studied classes of mobile genetic elements that harbour functional promoters, like Correia elements (Siddique et al. 2011), ERICs (De Gregorio et al. 2005) and REPIN (Bertels and Rainey 2011). But also examples of duplicated promoters not associated with known mobile elements (Usakin et al. 2005; Vandepoele et al. 2009) suggest that promoter reuse could represent a rampant and rapid mechanism of gene rewiring. In a recent publication Blount et al. (2012) identified a promoter capture event as a crucial step in the evolution of aerobic citrate utilization by a population of *Escherichia coli* in a long-term evolution experiment, and speculate that promoter capture may be an important and little appreciated adaptive force in genome evolution. Similarly, Bongers et al. (Bongers et al. 2003) described the activation of a silent lactate dehydrogenase gene by promoter recruitment in *Lactococcus lactis*. In both studies insertion sequences (IS) were involved in promoter mobility, though Blount et al. also found cases that were not associated with IS elements. In order to estimate the relevance of promoter recruitment in genome evolution we made a conservative inventory of such events in prokaryotes, that was recently published in *Nucleic Acids Research* (Matus-Garcia et al. 2012), chapter 5a of this thesis.

Tip of the Iceberg

To assess the extent of promoter reuse in bacteria we looked for groups of bacterial genes per genome that share highly similar sequences upstream of their transcriptional start site, but do not have obvious flanking paralogous coding sequences. More specifically, we extracted *in silico* the DNA region between positions -150 and -50 relative to the start of translation for all genes in a genome (including plasmids), except when this overlapped with the coding region or promoter region of a flanking gene. In *Escherichia coli* the majority of the transcriptional start sites were shown to be between 20 and 40 nucleotides

upstream of the translational start site (Mendoza-Vargas et al. 2009), so most of our upstream sequence fragments will not contain the important -10 (Pribnow) box, but should include the -35 element. Using BLAST (Altschul et al. 1997) we then searched for sequence pairs that matched with 80% or more nucleotide identity over at least 50 nucleotides, to select for highly similar regions rather than for short conserved DNA elements. Sequences with more than one hit in the database were clustered into families. Sequence pairs that in addition showed a high nucleotide identity in their adjacent coding sequences were assumed to be paralogs and excluded because for this study we were interested in the independent mobility of promoters, not duplicated regions (for details see the Materials and Methods in chapter 5a).

We analysed all available complete prokaryotic genomes (1,362; July 2011) and even with our strict selection criteria found over 4,000 families of highly similar sequences upstream of apparently unrelated coding sequences. The majority of these families actually consist of pairs that on average share 92% nucleotide identity, meaning that at least 46 out of 50 base pairs were conserved, but we also found pairs that were completely identical over 100 base pairs. Whether this level of high identity is the result of a strong selective pressure, or indicative of recent duplication events remains to be investigated. We termed these homologous non-coding sequences Putative Mobile Promoters, PMPs. In fact, some of these sequences likely are not promoters but have a different function that causes their conservation. Looking for known elements in our PMP set we actually found 42 tRNAs, 83 resembled other RNA families like the regulatory riboswitches, and interestingly 210 were known insertion sequences (Siguier et al. 2006). The > 4,000 families that our study uncovered represent only a small sample of a large pool of repeated DNA in promoter regions, a conservative reference of promoter reuse in prokaryotes. We anticipate many relevant examples of the phenomenon remain undetected because of our strict criteria. For example, filtering out paralogous genes also removes mobile promoters that extend into the coding region, like reported cases of Correia elements that overlap with an ORF (Siddique et al. 2011). In addition, our initial extraction of promoter sequences is sensitive to wrongly annotated translational start sites, which is a known issue with genome annotation pipelines (Nielsen and Krogh 2005).

Horizontal Promoter Transfer?

More surprising even than the large number of promoter pairs sharing high nucleotide identity within one bacterial genome is that about 6% are shared between distantly related species. Clustering these based on sequence similarity resulted in 62 distinct groups, of which four are present in species that are related only by belonging to the same phylum (Table 1). As expected, inter-taxon transfers seem to decrease with phylogenetic distance and at the domain level, i.e. between Archaea and Bacteria, no transfer events were observed. Some non-coding

sequence elements like tRNAs are very well conserved over large evolutionary distances (Saks and Conery 2007), but if highly similar sequences are found only in small number of distantly related species horizontal gene transfer is a more likely scenario. The large majority of the PMPs are located on a chromosome, but for one group of PMPs all members are in fact on plasmids. These plasmids are associated with multiple-drug resistance in pathogenic Salmonella (Fricke et al. 2009) and are frequently transferred between bacterial species.

Although the genetic code for translating DNA to protein is extremely well conserved between species as distant as *Escherichia coli* and *Homo sapiens*, transcriptional *cis*-regulatory elements are much more variable (Doniger and Fay 2007) and their activity can differ even between strains of the same species (Hendriksen et al. 2007; van Hijum et al. 2009). 3 It can therefore be expected that the 62 homologous PMPs are not primarily transcription factor binding sites, but rather have other (regulatory) functions causing their high conservation. Indeed, two of the PMPs that are shared between families of bacteria are known S-adenosylmethionine (SAM) binding riboswitches (Weinberg et al. 2007). The other 60 PMPs however did not resemble any of the RNA families included in the Rfam database (Gardner et al. 2011), so their function at present remains uncovered.

Branch point	**Count**
Genera	28
Families	12
Orders	9
Class	9
Phylum	4
Domain	0
	62

Table 1. Number of PMPs shared by bacterial genomes from different genera, families, orders, etc.

We conclude that we have uncovered a large number of putative mobile promoter families, present in numerous bacterial genomes. These may be involved in rapid adaptive processes via transcriptional rewiring, or include post-transcriptional regulatory functions. The ways these PMPs move within and between genomes is still unknown, but due to the large number of families, this may include diverse mobilization mechanisms.

Finally, although transcription regulation in eukaryotes is more complex than in bacteria, it seems obvious that also in eukaryotes promoter reuse offers a

mechanism for rapid adaptation of gene expression. It would therefore be very interesting to extent our analysis to this domain, especially now more genomes and transcriptomes are becoming available that greatly facilitate the mapping of the core promoters.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bertels F, Rainey PB. 2011. Within-genome evolution of REPINs: a new family of miniature mobile DNA in bacteria. *PLoS Genet* **7**: e1002132.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**: 513–518.

Bongers RS, Hoefnagel MHN, Starrenburg MJC, Siemerink MAJ, Arends JGA, Hugenholtz J, Kleerebezem M. 2003. IS981-mediated adaptive evolution recovers lactate production by ldhB transcription activation in a lactate dehydrogenase-deficient strain of Lactococcus lactis. *J Bacteriol* **185**: 4499–4507.

De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP. 2005. Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J Bacteriol* **187**: 7945–7954.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* **3**: e99.

Fricke WF, Welch TJ, McDermott PF, Mammel MK, LeClerc JE, White DG, Cebula TA, Ravel J. 2009. Comparative genomics of the IncA/C multidrug resistance plasmid family. *J Bacteriol* **191**: 4750–4757.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, et al. 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res* **39**: D141–5.

Hendriksen WT, Silva N, Bootsma HJ, Blue CE, Paterson GK, Kerr AR, de Jong A, Kuipers OP, Hermans PWM, Mitchell TJ. 2007. Regulation of gene expression in Streptococcus pneumoniae by response regulator 09 is strain dependent. *J Bacteriol* **189**: 1382–1389.

Matus-Garcia M, Nijveen H, van Passel MWJ. 2012. Promoter propagation in prokaryotes. *Nucleic Acids Res* **40**: 10032–10040.

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, et al. 2009. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS ONE* **4**: e7526.

Nielsen P, Krogh A. 2005. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21**: 4322–4329.

Saks ME, Conery JS. 2007. Anticodon-dependent conservation of bacterial tRNA gene sequences. *RNA* **13**: 651–660.

Siddique A, Buisine N, Chalmers R. 2011. The transposon-like Correia elements encode numerous strong promoters and provide a potential new mechanism for phase variation in the meningococcus. *PLoS Genet* **7**: e1001277.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**: D32–6.

Usakin LA, Kogan GL, Kalmykova AI, Gvozdev VA. 2005. An alien promoter capture as a primary step of the evolution of testes-expressed repeats in the Drosophila melanogaster genome. *Mol Biol Evol* **22**: 1555–1560.

van Hijum SAFT, Medema MH, Kuipers OP. 2009. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol Mol Biol Rev* **73**: 481–509– Table of Contents.

van Passel MWJ, Marri PR, Ochman H. 2008. The emergence and fate of horizontally acquired genes in Escherichia coli. *PLoS Comput Biol* **4**: e1000059.

Vandepoele K, Andries V, van Roy F. 2009. The NBPF1 promoter has been recruited from the unrelated EVI5 gene before simian radiation. *Mol Biol Evol* **26**: 1321–1332.

Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, et al. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res* **35**: 4809–4819.

Chapter 6

General discussion

A New Biology

The nature of biological research is profoundly changing now the parts of biological systems have been extensively catalogued. Initiatives like the Human Genome Project (Collins et al. 2003) dramatically accelerated the development of new high-throughput measurement technologies. The current level of insight into the building blocks of life not only allows biologists to address questions that were previously out of reach, it also stimulates approaches to study biology at different organisational levels. These more holistic approaches acknowledge that a gene does not work in isolation, but functions in concert with the other genes in the organism, and that the organism in turn interacts with other organisms and environmental factors in the ecosystem (Woese 2004). This multi-level or holistic view on biology offers a broad range of new scientific challenges that have attracted, besides biologists, physicists, mathematicians, chemists, computer scientists and engineers. Both the National Academy of Sciences in the United States and the Royal Netherlands Academy of Arts and Sciences have recognised the potential of this multi-disciplinary field, termed New Biology, in addressing challenging scientific and societal problems. They have formulated recommendations to best capitalize on the benefits of the New Biology. Bioinformatics has an important role in this, as "information is the fundamental currency of the New Biology" (USA: National Research Council, 2009). Especially dealing with large volumes of data ('big data') that are produced by the high-throughput technologies is a new experience for current biological researchers and not something they were properly trained in (Pevzner and Shamir 2009).

The main challenge of working with big data is to process it in such a way that it is suitable for human inspection, and biological knowledge can be gained. While gene expression measurements for a single gene could still be analysed more or less by hand (Doerwald et al. 2001), this is not practical anymore when the expression of all genes in the genome is measured. Firstly, because the measurement itself produces data in quantities that cannot be captured manually, and secondly because the subsequent analysis and comparison of tens of thousands of expression values requires computer assistance, for instance to find clusters of genes showing similar expression patterns (Trapnell et al. 2012). And it is in this phase of a typical big data research project that bioinformatics has an important role to play. Usually the data from the measuring device first has to be mapped to biological entities like genes, metabolites, proteins, etc. Then further statistical analysis, clustering, or high-level visualisation can be performed to transform or summarise the data to a level that it is suitable for biological interpretation. This last step is often a cyclic process, in which conclusions from one data filtering, clustering or statistical analysis step lead to new questions that can be addressed with an alternative analysis producing a different view of the data, a process that is sometimes referred to as 'exploratory data analysis' (Tukey 1977), or 'playing with the data'. However, usually this is not as easy as the last expression suggests,

and requires bioinformatics experience and specialist software tools.

I identify three different approaches for successful incorporation of bioinformatics analysis in research projects:

- Empowerment of the biologist with user-friendly software tools
- Collaboration of biologists and bioinformaticians (team work)
- Improving the computational skills of the biologist

Examples of the first two approaches were given in the previous chapters. I developed user-friendly software tools that can be used directly by bench biologists for their research (chapters 2 and 3), and I collaborated with biologists on projects that required custom bioinformatics analyses to address a biological question (chapters 4, 5a and 5b). The third approach calls for a shift in focus for biology education, not only at the bachelor level, but also at the master and PhD levels. Just like molecular biology laboratory techniques, basic computational biology skills training should be part of an undergraduate biology program. From my personal experience as a bioinformatics teacher, learning some basic programming skills already unlocks powerful new ways for researchers to address biological questions. I do not foresee that any of the three approaches will be dominant in the near future, since each has its merits and shortcomings. I will discuss the three approaches individually, and in the process present a perspective of how biology could profit best from the new technological opportunities.

Empowerment

Putting a strong focus on user interaction in the design of software for biological data analysis can make it much more accessible to researchers without expert computer skills. With powerful user-friendly software tools the biologist can intuitively examine his or her experimental results, while the informatics aspects of the software remain largely hidden, i.e. are 'black-boxed'. Now, a lot of progress has already been made generally in the way humans interact with computers: in the early days computer users still had to have in depth knowledge of the internal workings of a computer, and nowadays even a toddler can work (play) with computer devices like Apple's iPad. However, for scientific analysis user-friendly tools are still scarce, mostly because of the complexity and heterogeneity of scientific data and analytical tasks. For the software that is available we can make a distinction between multi-purpose tools and tools that perform a specific function.

In the category of multi-purpose, 'Swiss Army Knife' type of software that is used by biologists, there are mainly commercial software packages. A powerful software tool for data analysis that was widely adopted by biologists is undeniably the spreadsheet. With the use of a spreadsheet numerical data can be ordered, normalized, statistically analysed, summarized and plotted in various graphs.

Popular spreadsheet programs like Excel (Microsoft) and Calc (OpenOffice/ LibreOffice) are relatively easy to use, are readily available to most scientists, and can handle reasonably large volumes of data efficiently. Largely lacking in these spreadsheet programs is the biological context and the availability of bioinformatics algorithms. Software packages like Spotfire® (Ahlberg 1996) and Agilent's GeneSpring® are more tailored to the bioinformatics domain and are popular for analysis of, for instance, microarray data. For next generation sequencing data analysis the software workbenches of the Danish company CLCbio are widely used. Although access might be limited by the price, most commercial software packages are relatively easy to use. Spotfire® was for instance designed with guidance of Ben Shneiderman, an authority in the field of information visualisation (Saraiya et al. 2004). Moreover, commercially developed software has the potential to be of higher quality in terms of correctness, stability, efficiency and documentation. Published analyses performed using these tools can also be readily repeated by other groups, provided they have a license. Still these packages cover only a limited area of well-established analysis methods that have a large potential user group. Writing professional software is laborious and expensive, and therefore economically not very interesting for a niche application with mainly underfunded academic users.

Next to these commercial multi-purpose data analysis tools, there is a strong need for more specialist tools aimed at bench biologists. As a response to this the number of user-friendly software tools that perform a limited set of tasks is steadily growing. These applications are often developed by members of the scientific community in open source projects. An example of such a tool is Primer3Plus, described in chapter 2, a tool developed to help wet lab biologists with the task of designing primers. By providing it as a publicly available web interface it has attracted a large number of users (>100,000 counting unique internet addresses) and was cited over 400 times. Designing primers is a relatively simple task that does not require in depth analysis, and although the underlying algorithm (Primer3) has many configurable parameters, the default settings are usually good enough. From a usability point of view the used design method and the resulting user interface may not be up to the latest standards, but apparently this does not invalidate the application. Association with the already popular Primer3 software, providing credibility, and the user-centric design of the interface can explain the popularity of the tool. A similar project that was recently completed is the QualitySNPng application (chapter 3). This builds upon the QualitySNP pipeline that was published in 2006 and that had to be used from the command line. QualitySNPng was designed for non-computer savvy biologists working with next generation sequencing data. Through a graphic user interface they can interact with the software to tune the SNP detection routine based on their knowledge of the biological system. For instance the expected frequency of the minor allele of a SNP strongly depends on the ploidy of the organism under study.

By looking in detail at the sequence reads and the discovered SNPs, the biologist can decide to redo the SNP detection with stricter settings to reduce unreliable SNP calls.

The enthusiasm with which the two above described software tools were received by the user community and their acceptances for publication in 'Nucleic Acids Research' illustrates their value to the biological community.

Team work

Two steps can be discerned when working with big data: in the first step the more or less raw data from the measuring device are mapped to biological entities and in the second step the data are reduced or summarised to a level that allows biological interpretation. For standard analyses these steps can be encapsulated in a software tool that lab scientists can use without special computer training. Examples of such tools were given in the previous section. In contrast, if the experiment calls for a more exploratory data analysis, such a software tool probably will not suffice (if it even exists), and expertise of the applicable methods for data summarisation or reduction is needed. Just providing the biologist with an intuitive interface to interact with the software does not help if he or she does not know the characteristics of the offered analysis methods and may even lead to wrong conclusions. Such cases with non-standard or novel analysis approaches will benefit from a close collaboration between a biologist and a bioinformatician.

In such a collaborative project both experts have complementary roles. The biological question is clear in the mind of the biologist, including knowledge of the relevant literature and some idea of how to address the question. The bioinformatician can translate the question to a software analysis pipeline and format the results in a biologically meaningful way for the biologist to evaluate. The bioinformatician should know enough biology to understand the biological question and to communicate at a high level with the biologist. Ideally the overlap in knowledge and background ensures efficient interaction of ideas and leverages the unique knowledge and skills that each of the team members possesses to come to the desired biological analysis. Depending on the specificity of the biological question, these projects can range from a straightforward translation of the biological question into a technical approach, to a more explorative data analysis process involving several cycles of analysis and biological interpretation. The role of the bioinformatician can vary likewise between mostly technical and service oriented, to a more scientific one.

The Human Short Peptide variation database as described in chapter 4 is an example of a software tool that was developed together with biologists. First a specialist data source was developed for mapping captured mass spectra from the detection device to their corresponding biological entities, proteins.

Several protocols for creating the database were tried before the biologists were satisfied with the number and relevance of the identified proteins. Subsequently the analysis steps that were at first performed manually by the biologists were automated as much as possible by data integration. By allowing easy interaction through a web interface the functionality of the database was made available to a wider group of users and the involved biologists now use it frequently in their current research without the need of bioinformatics support. Although use by other biologists is still limited, probably due to the specialist nature of the database, it shows how a collaborative project can produce a software tool that supports the above described empowerment approach.

An example of a more exploratory data analysis process can be found in chapters 5a and 5b, which describe a project to assess the genomic mobility of bacterial promoters. The used software tools were mainly custom created scripts, BLAST (Altschul et al. 1997) and publicly available databases. Because the biological question was quite broad, it required many iterations of analysis followed by data interpretation before a conclusion could be reported. Some initial assumptions had to be reconsidered and several misconceptions between the biologist and the bioinformatician took some time to surface, but in general communication worked well and even led to knowledge transfer. Although the scripts could not be thoroughly tested for lack of a large manually validated data set, the results were critically evaluated for their biological plausibility and a number of interesting results were repeated with careful inspection of the outcomes of the intermediate steps. The counts and sizes of the mobile promoter families are currently being used to model the dynamics of promoter duplication with the aid of a mathematician.

Both collaborative efforts described above were initiated by a biological question requiring computational analyses for which no off the shelf tools existed. Only the results were published, as a database with an accompanying web interface and in publications. The developed scripts were not made publicly available, because they were specifically created for a single application and as such were not thoroughly tested to work well with all kinds of different input data.

Training biologists in computational methods

In the third approach biologists are empowered with training in computational methods, effectively putting the biologist and the bioinformatician in the same head (Maclean and Kamoun 2012). When the biologist is proficient with computers he or she can combine biological insight to ask the right questions with the computer skills to address these questions. This avoids miscommunication and can lead to creative use of computational techniques. Many bench biologists have already taken up the challenge to learn how to program custom software and thus do the analyses they would like to perform.

Traditionally bioinformatics software tools were created to work through a (Linux) command-line interface rather than via a user-friendly graphical user interface and this has only recently started to change. Additionally, most tools only perform one specialist function (Rice et al. 2000), require specific pre-processed input and produce their output in one of the many, often redundant, data formats. For a complete bioinformatics analysis usually several tools have to be used sequentially in a bioinformatics pipeline or workflow. Construction of these pipelines can be severely complicated by the 'environment of creative chaos' (Stein 2002) that exists in the realm of bioinformatics software and data formats. Using custom scripts to reformat the output of one program to acceptable input for another is a common task. Linux command-line use and experience with a scripting language like Python or Perl are then indispensible.

My personal experience from the Advanced Bioinformatics MSc. course taught at the Wageningen University has shown that students without any previous programming or command line experience can learn to build a simple bioinformatics analysis pipeline with about two weeks of training. But I also observed that there are huge variations in the aptitude students have for programming, and for a substantial number of them it will require an extensive training effort before they are able to creatively address biological questions with software tools written by themselves. Some even argue that many people will never learn to program (Dehnadi and Bornat 2006). This variation in the ability to acquire programming skills is reflected in the field, where some biologists still require bioinformatics assistance for relatively simple data analysis tasks or avoid computational analyses altogether, while others have rapidly adapted to make full use of the new opportunities of big data and now spend most of their time doing bioinformatics analyses. In a sense the latter category of researchers could be called 'bioinformaticians', although they usually lack formal training in computer science and bioinformatics algorithms. But if that were a requirement for being a bioinformatician, then only a small minority should be allowed that title.

Bioinformatics education is slowly making its way into many undergraduate and graduate biology programs (Honts 2003; Cummings and Temple 2010; Maloney et al. 2010). Interesting initiatives like the Rosalind platform (http://rosalind.info) offer a stimulating environment for learning online how to solve bioinformatics problems with computer programming. The Bioconductor project (Gentleman et al. 2004) is an example of a successful computational resource for biologists who know how to program in the R statistical programming language.

But there is hope for biologists who do not know how to write computer programs. The cause of this is the growing popularity of workflow systems that allow users

to create bioinformatics analysis pipelines by connecting existing tools through a graphical user interface. Examples of these systems are e-BioFlow (Wassink et al. 2008), Taverna (Hull et al. 2006; Wolstencroft et al. 2013), and Galaxy (Goecks et al. 2010). For next generation sequencing analysis Galaxy combines an easy to use web interface for biologists with a powerful and flexible cloud based computing infrastructure (Afgan et al. 2011). There are many advantages of workflow tools over custom programmed pipelines: they largely remove the need to write scripts, offer sophisticated flow control and provide readily sharable analysis protocols (Goble et al. 2010). More complicated aspects of workflow systems that require computer skills, and thus limit their usability by biologists, include writing wrappers for making new tools available as workflow components and properly connecting the result of one component as the input of another component. Also managing efficient data transport through a workflow can be a challenge, especially when gigabytes of data are transported between distributed analysis components. Moreover, the user still has to have a good idea of the characteristics of the different tools in the workflow protocol.

Learning and maintaining bioinformatics skills gives the biologist the capacity to design custom experiments that best match the biological question to the available data analysis methods, although it also means that the researcher has to divide his or her time between keeping up to date with the biological topic and novel bioinformatics algorithms, software packages and databases. The biologist ideally gets frequent feedback from local expert bioinformaticians, statisticians and software developers for validation of chosen analysis approaches, statistical assumptions and written software code. To facilitate this, regular multidisciplinary work discussion and journal club sessions can help (Maclean and Kamoun 2012).

Perspective: evolution of biological research

Next to the above addressed need to aid current wet lab biologist with applying computation analyses in their projects, it is clear that biological research in general has to adapt to its changing environment. With the developing technology the way we can study biological systems is also changing, as we move for instance from examining a single gene to analysis at the organism or population level with genomics, transcriptomics, proteomics and metabolomics data. The field of systems biology that aims to model biological systems is rapidly growing (Kitano 2002). New biological insights can be gained by studying complete biological systems, for instance the discovery of recurring motifs in transcription regulation networks (Alon 2007), regulatory motifs that are similar to those used in engineering. New statistical challenges arise when assessing the significance of effects in large data volumes (Benjamini and Hochberg 1995). To make full use of the new possibilities in data collection, the biologist will have to be educated in available data analysis methods, perhaps not to the level of being an expert, but at least enough to understand how and when to apply them (Ditty et al. 2010). User-

friendly software can hide away or black-box the technical implementation of an analysis method, but also the empowered biologist should have a basic idea of what an algorithm does to be able to properly interpret its results. The same holds true for the use of databases, some sense of the reliability and completeness of the data is essential to avoid drawing wrong conclusions. Every biologists working with high throughput data will need a basic understanding of computational methods for data analysis.

Expert bioinformaticians and bioinformatics assistants have three important roles to play in the transition of biology into a big data science. The first role is that of teacher, educating biologists and training students in bioinformatics. The second role is in assisting biologists with analysing their data by building custom analysis pipelines and writing dedicated analysis scripts. This supportive role should also include a proper scientific computing infrastructure, with high speed computer clusters, but also facilities like version control, software development frameworks and data management and provenance. The third role is more fundamental, developing new analysis methods and applications. The role can change during a project, for instance when a collaboration project requires the development of a new algorithm, or user training is organized for a newly created software tool. Each role requires different skills and expertise. State of the art software engineering practices (Baxter et al. 2006) should be applied in developing a new software tool, whereas collaboration with a biologist on a project requires good communication skills and enough biological knowledge to understand the biological problem. Bioinformaticians have an important part to play in shaping a new data rich biology.

The distinction between biology and bioinformatics as separate fields will be more and more difficult to make as biology gets rapidly computerized to make full use of the 'big data' technologies (Ouzounis 2012), similar to what happened with the field of molecular biology (Morange 2008) when researchers from other biological disciplines saw the potential of molecular techniques for studying biological systems. A clear sign that bioinformatics is gaining interest among biologists, is the rapidly growing popularity of the bioinformatics MSc and PhD courses taught at Wageningen University. To optimally prepare current and future biology students for the data intensive field they will work in, BSc, MSc and PhD educational programs should have a strong computational biology component (Pevzner and Shamir 2009). At the same time bioinformatics software development needs to be professionalised with a strong focus on powerful user interaction (Kumar and Dudley 2007; Pavelin et al. 2012) and efficient processing of large data volumes (Trelles et al. 2011). Our success in adapting to the new big data reality will be the main factor determining how well we are able to leverage this information to start understanding biology at the systems level (Kanehisa and Bork 2003).

REFERENCES

Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, Nekrutenko A, Taylor J. 2011. Harnessing cloud computing with Galaxy Cloud. *Nat Biotechnol* **29**(11): 972-974.

Ahlberg C. 1996. Spotfire: an information exploration environment. *ACM SIGMOD Record* **25**(4): 25-29.

Alon U. 2007. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**(6): 450-461.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Baxter SM, Day SW, Fetrow JS, Reisinger SJ. 2006. Scientific software development is not an oxymoron. *PLoS Comput Biol* **2**(9): e87.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodo* **57**(1): 289-300.

Collins FS, Morgan M, Patrinos A. 2003. The Human Genome Project: lessons from large-scale biology. *Science* **300**(5617): 286-290.

Cummings MP, Temple GG. 2010. Broader incorporation of bioinformatics in education: opportunities and challenges. *Brief Bioinform* **11**(6): 537-543.

Dehnadi S, Bornat R. 2006. The camel has two humps (working title). *Middlesex University*.

Ditty JL, Kvaal CA, Goodner B, Freyermuth SK, Bailey C, Britton RA, Gordon SG, Heinhorst S, Reed K, Xu Z et al. 2010. Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol* **8**(8): e1000448.

Doerwald L, Nijveen H, Civil A, van Genesen ST, Lubsen NH. 2001. Regulatory elements in the rat betaB2-crystallin promoter. *Exp Eye Res* **73**(5): 703-710.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**(10): R80.

Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P et al. 2010. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res* **38**(Web Server issue): W677-682.

Goecks J, Nekrutenko A, Taylor J, The Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**(8): R86.

Honts JE. 2003. Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biol Educ* **2**(4): 233-247.

Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. 2006. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* **34**(Web Server issue): W729-732.

Kanehisa M, Bork P. 2003. Bioinformatics in the post-sequence era. *Nature genetics* **33 Suppl**: 305-310.

Kitano H. 2002. Systems biology: a brief overview. *Science* **295**(5560): 1662-1664.

Kumar S, Dudley J. 2007. Bioinformatics software for biologists in the genomics era. *Bioinformatics* **23**(14): 1713-1717.

Maclean D, Kamoun S. 2012. Big data in small places. *Nat Biotechnol* **30**(1): 33-34.

Maloney M, Parker J, Leblanc M, Woodard CT, Glackin M, Hanrahan M. 2010. Bioinformatics and the undergraduate curriculum essay. *CBE Life Sci Educ* **9**(3): 172-174.

Morange M. 2008. The death of molecular biology? *Hist Philos Life Sci* **30**(1): 31-42.

Ouzounis CA. 2012. Rise and demise of bioinformatics? Promise and progress. *PLoS Comput Biol* **8**(4): e1002487.

Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G, Steinbeck C. 2012. Bioinformatics meets user-centred design: a perspective. *PLoS Comput Biol* **8**(7): e1002554.

Pevzner P, Shamir R. 2009. Computing has changed biology-biology education must catch up. *Science* **325**(5940): 541-542.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**(6): 276-277.

Saraiya P, North C, Duca K. 2004. An evaluation of microarray visualization tools for biological insight. *Proc of IEEE Symposium on Information Visualization 2004*: 1-8.

Stein L. 2002. Creating a bioinformatics nation. *Nature* **417**(6885): 119-120.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Prot* **7**(3): 562-578.

Trelles O, Prins P, Snir M, Jansen RC. 2011. Big data, but are we ready? *Nat Rev Genet* **12**(3): 224.

Tukey JW. 1977. *Exploratory data analysis*. Pearson.

Wassink I, Rauwerda H, van der Vet P, Breit T. 2008. e-BioFlow: Different perspectives on scientific workflows. In *2nd Internationall Conference on Bioinformatics Research and Development (BIRD)*, (ed. MK Elloumi, J.; Linial, M.; Murphy, R.; Schneider, K.; Toma, C.), pp. 243-257, Vienna, Austria.

Woese CR. 2004. A new biology for a new century. *Microbiol Mol Biol Rev* **68**(2): 173-186.

Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P et al. 2013. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41**(Web Server issue): W557-561.

Summary

Biology is becoming a data-rich science driven by the development of high-throughput technologies like next-generation DNA sequencing. This is fundamentally changing biological research. The genome sequences of many species are becoming available, as well as the genetic variation within a species, and the activity of the genes in a genome under various conditions. With the opportunities that these new technologies offer, comes the challenge to effectively deal with the large volumes of data that they produce. Bioinformaticians have an important role to play in organising and analysing this data to extract biological information and gain knowledge. Also for experimental biologists computers have become essential tools. This has created a strong need for software applications aimed at biological research. The chapters in this thesis detail my contributions to this area. Together with molecular biologists, plant breeders, immunologists, and microbiologists, I have developed several software tools and performed computational analyses to study biological questions.

Chapter 2 is about Primer3Plus, a web tool that helps biologists to design DNA primers for their experiments. These primers are typically short stretches of DNA (~20 nucleotides) that direct the DNA replication machinery to copy a selected region of a DNA molecule. The specificity of a primer is determined by several chemical and physical properties and therefore designing good primers is best done with the help of a computer program. Primer3Plus offers a user-friendly task-oriented web interface to the popular primer3 primer design program. Primer3Plus clearly fulfils a need in the biological research community as already over 400 scientific articles have cited the Primer3Plus publication.

Single nucleotide differences or polymorphisms (SNPs) that are present within a species can be used as markers to link phenotypic observations to locations on the genome. Chapter 3 discusses QualitySNPng, which is a stand-alone software tool for finding SNPs in high-throughput sequencing data. QualitySNPng was inspired by the QualitySNP pipeline for SNP detection that was published in 2006 and it uses similar filtering criteria to distinguish SNPs from technical artefacts like sequence read errors. In addition, the SNPs are used to predict haplotypes. QualitySNPng has a graphical user interface that allows the user to run the SNP detection and evaluate the results. It has already been successfully used in several projects on marker detection for plant breeding.

Single nucleotide polymorphisms can lead to single amino acid changes in protein sequences. These single amino acid polymorphisms (SAPs) play a key role in graft-versus-host (GVH) effects that often accompany tissue transplantations. A beneficial variant of GVH is the graft-versus-leukaemia (GVL) effect that is

sometimes witnessed after bone marrow transplantation in leukaemia patients. When the GVL effect occurs, the donor's immune cells actively destroy residual tumour cells in the patient. The GVL effect can already be elicited by a single amino acid difference between the patient and the donor. Currently, a small number of SAPs that can elicit a GVL effect are known and these are used to select the right bone marrow donor for a leukaemia patient. Together with researchers at the Leiden University Medical Center I developed a database to aid in the discovery of more such SAPs. We called this database the "Human Short Peptide Variation database" or HSPVdb. It is described in chapter 4.

The work described in chapter 5 is focused on the regions in bacterial genomes that are involved in gene regulation, the promoters. Intrigued by anecdotal evidence that duplication of bacterial promoters can activate or silence genes, we investigated how often promoter duplication occurs in bacterial genomes. Using the large number of bacterial genomes that are currently available, we looked for clusters of highly similar promoter regions. Since duplication assumes some sort of mobility, we termed the duplicated promoters: putative mobile promoters or PMPs. We found over 4,000 clusters of PMPs in 1043 genomes. Most of the clusters consisted of two members, indicating a single duplication event, but we also found much larger clusters of PMPs within some genomes. A number of PMPs are present in multiple species, even in very distantly related bacterial species, suggesting perhaps that these were subjected to horizontal gene transfer. The mobile promoters could play an important role in the rapid rewiring of gene regulatory networks.

Chapter 6 discusses how current biological research can adapt to make full use of the opportunities offered by the high-throughput technologies by following three different approaches. The first approach empowers the biologists with user-friendly software that allows him to analyse the large volumes of genome scale data without requiring expert computer skills. In the second approach the biologist teams up with a bioinformatician to combine in-depth biological knowledge with expert computational skills. The third approach combines the biologist and the bioinformatician in one person by teaching the biologist computational skills. Each of these three approaches has it merits and shortcomings, so I do not expect any of them to become dominant in the near future. Looking further ahead, it seems inevitable that any biologist will have to learn at least the basics of computational methods and that this should be an integral part of biology education. Bioinformatics might in time cease to exist as a separate field and instead become an intrinsic aspect of most biological research disciplines.

Samenvatting

Recente technologische ontwikkelingen hebben de analyse van biologische bouwstenen als DNA en eiwitten enorm verbeterd. Hierdoor kunnen onderzoekers biologische systemen op een tot voor kort ondenkbaar detailniveau bestuderen. De grote hoeveelheid aan meetgegevens die hierdoor beschikbaar komt maakt van het biologisch onderzoek in snel tempo een data-intensieve wetenschap. De uitdaging waar biologen nu voor staan is het vinden van bruikbare informatie in deze data-explosie. De computer is hierbij een essentieel gereedschap. Computationele analyse van biologische data is het vakgebied van de bioinformatica en de bioinformaticus is dus bij uitstek toegerust voor de data-intensieve biologie. Er is een sterke behoefte ontstaan aan softwaretoepassingen gericht op biologisch onderzoek. De hoofdstukken in dit proefschrift beschrijven mijn bijdragen op dit gebied. Tezamen met moleculair biologen, plantenveredelaars, immunologen en microbiologen heb ik verschillende softwaretoepassingen ontwikkeld en ik heb computationele analyses uitgevoerd om biologische vraagstukken te bestuderen.

Hoofdstuk 2 behandelt Primer3Plus, een webapplicatie waarmee biologen DNA-primers kunnen ontwerpen. Primers zijn korte stukjes DNA (±20 nucleotiden) die gebruikt worden voor het vermenigvuldigen van een specifiek deel van een DNA molecuul om het zo beter te kunnen bestuderen. De specificiteit van een primer wordt bepaald door verschillende chemische en fysische eigenschappen, en het ontwerpen van goede primers gebeurt met behulp van een computerprogramma. Primer3Plus laat de gebruiker kiezen tussen verschillende standaardtoepassingen voor primers en toont alleen de invoervelden die van belang zijn voor de gekozen toepassing. Het voorziet in een behoefte in het biologisch onderzoeksveld, zoals blijkt uit de meer dan 400 wetenschappelijke artikelen die de Primer3Plus publicatie citeren.

Hoofdstuk 3 beschrijft QualitySNPng, een softwareapplicatie die de bioloog helpt om kleine verschillen in het DNA te vinden. Een verschil van één enkele letter in hetzelfde gen van twee verschillende individuen van hetzelfde soort wordt een SNP genoemd, kort voor “Single Nucleotide Polymorphism”. Deze SNPs kunnen bijvoorbeeld gebruikt worden als merkers om fenotypische observaties te koppelen aan specifieke genen. QualitySNPng is geïnspireerd op de QualitySNP pipeline voor SNP detectie die in 2006 is gepubliceerd, en het gebruikt vergelijkbare filtercriteria om SNPs te onderscheiden van technische artefacten zoals sequentieleesfouten. QualitySNPng heeft een gebruikersvriendelijke interface waarmee de gebruiker de SNP detectie kan uitvoeren en de resultaten kan inspecteren. De software is al succesvol toegepast in verschillende projecten voor merkerdetectie in plantenveredeling.

Omdat het DNA de bouwplannen voor de eiwitten van het organisme bevat, kunnen verschillen in het DNA leiden tot veranderingen in de manier waarop de bouwstenen van een eiwit, de aminozuren, aan elkaar gekoppeld zijn. Dit maakt SNPs ook interessant voor de immunologie, en dan specifiek in verband met weefseltransplantaties. Bepaalde verschillen in het DNA van de donor ten opzichte van het DNA van de patiënt kunnen ertoe leiden dat eiwitten van de donor door het immuunsysteem van de patiënt als lichaamsvreemd worden gezien. Dit kan een heftige afweerreactie uitlokken in de patiënt tegen het getransplanteerde weefsel van de donor, het zogenaamde graft-versus-host (GVH) effect. Een gunstige variant van dit GVH is het graft-versus-leukemia (GVL) effect dat soms optreedt na een beenmergtransplantatie bij patiënten met leukemie. Bij het GVL effect wordt het immuunsysteem van de donor gebruikt om tumorcellen in de patiënt op te ruimen. Dit kan dan het terugkeren van de leukemie voorkomen. Het GVL effect kan al worden uitgelokt door één enkel aminozuurverschil tussen patiënt en donor. Er zijn maar een klein aantal aminozuurverschillen bekend die het GVL effect kunnen veroorzaken, en deze worden gebruikt om de beste beenmergdonor voor een leukemiepatiënt te selecteren. Samen met onderzoekers van het Leids Universitair Medisch Centrum heb ik een databank ontwikkeld om te helpen met het ontdekken van meer van deze aminozuurverschillen. Deze databank hebben we de “Human Short Peptide Variation database” of HSPVdb genoemd. De HSPVdb staat beschreven in hoofdstuk 4.

Het werk in hoofdstuk 5 gaat over de gebieden in bacteriële genomen die betrokken zijn bij genregulatie, de promotors. Uit de literatuur is bekend dat duplicatie van bacteriële promotors genen kan activeren of juist deactiveren. Wij hebben uitgezocht hoe vaak promotorduplicatie optreedt in bacteriële genomen. Gebruikmakend van het grote aantal bacteriële genomen dat tegenwoordig beschikbaar is, zochten we naar groepen sterk gelijkende promotoren. Omdat duplicatie van een stukje DNA een vorm van mobiliteit veronderstelt, hebben we de gedupliceerde promotors “Putative Mobile Promoters” genoemd, of PMP’s. We hebben meer dan 4.000 PMP groepen gevonden in 1.043 verschillende bacteriële genomen. De meeste van deze groepen bestaan uit twee leden, wijzend op een enkele duplicatie, maar we vonden ook veel grotere groepen van PMP’s in sommige genomen. We vonden ook PMP’s voorkomend in meerdere soorten, zelfs in enkele evolutionair gezien ver verwijderde bacteriële soorten. Dit zou erop kunnen wijzen dat bacteriën van verschillende soorten deze promotors aan elkaar hebben doorgegeven door middel van een proces dat bekend staat als horizontale genoverdracht. Mobiele promotors kunnen een belangrijke rol spelen in het snel aanpassen van regulatoire genennetwerken, bijvoorbeeld als een bacteriepopulatie zich moet aanpassen aan een veranderende omgeving.

In hoofdstuk 6 ten slotte schets ik een perspectief voor efficiënt gebruik in het biologisch onderzoek van de grote hoeveelheden gegevens die met de

nieuwe technologieën worden gegenereerd. De nadruk ligt hierbij op de veelal ontbrekende expertise in computationele data-analyse bij experimenteel biologen. Ik identificeer drie verschillende benaderingen. Bij de eerste benadering neemt gebruikersvriendelijke software de technische details van de data-analyse grotendeels uit handen van de bioloog, waardoor hij grote hoeveelheden biologische data kan analyseren zonder te beschikken over uitgebreide computerexpertise. In de tweede benadering werkt de bioloog samen met een bioinformaticus, om zo biologische kennis te combineren met computerexpertise. De derde benadering combineert de bioloog en de bioinformaticus in één persoon door de bioloog te trainen in computationele vaardigheden. Welke benadering het meest geschikt is zal per project verschillen. Onafhankelijk van de gekozen benadering zal de onderzoeker wel een algemeen begrip van de gebruikte analysemethode moeten hebben om de resultaten goed te kunnen duiden. De conclusie dat computationele data-analyse een vaste plaats in de biologieopleiding verdient ligt dan ook voor de hand. Bioinformatica zal een intrinsiek onderdeel worden van de verschillende biologische onderzoeksvelden.

Dankwoord

Al in 2005 tijdens mijn sollicitatiegesprek met Jack Leunissen voor de functie van wetenschappelijk programmeur in zijn leerstoelgroep kwam het onderwerp ‘proefschrift’ op tafel. De jaren daarna bleef het als actiepunt op de agenda staan, maar zonder de nodige urgentie. Dat werd anders toen Jack in 2011 ziek werd. Helaas heeft Jack het voltooien van dit boekje niet meer meegemaakt, hij overleed op 14 mei 2012. Zijn bijdrage aan dit proefschrift is groot.

Paul, als coach heb je het afgelopen anderhalf jaar de juiste vragen gesteld en aanwijzingen gegeven om de verschillende hoofdstukken samen te smeden tot dit proefschrift. Het is altijd een genoegen om je in Wageningen te ontvangen.

Ton, dank voor je vertrouwen en steun.

Veel dank ben ik ook verschuldigd aan de medeauteurs van de verschillende hoofdstukken: Mark, Peter, Mariana, Michel, Ben, Andreas, Xiangyu, René, Martijn, Danny, Brechtje, Aurélie, Machiel, Fred, Arnoud en Chopie.

Douwe, Maria en Marie-José, hartelijk dank voor de onmisbare hulp bij de verschillende organisatorische en regelactiviteiten.

Mijn kamergenotes Sandra en Judith, jullie hebben de laatste loodjes een stukje lichter gemaakt.

En dan nog deze (incomplete en ongesorteerde) lijst met mensen die de afgelopen jaren hebben gezorgd voor de nodige inspiratie, motivatie en dieverdoatsie:
Jan, HJ, Ernest, Blaise, Pieter, Pjotr, Ernst, Ke, Heleen, Anand, Adrien, Audrey, Patrick, Arnold, Erik, Klaas-Jan, Dirk, Job, AJ, Edouard, Pierre, Erwin, Bernd, Philip, Guido, Thomas, Roeland, Koen, Mark, Tom, Jose, Sander, Sven, Yiannis, Olga, Riccardo, Ole, Nital, Lenie, Vincent, Lars, Joachim, Felippe, Arwa, Saulo, Lettie, Siebe, Gerben, Eric, Marc, Julian, Koos, Ine, Freek, David, Linda, Arjen, Jifeng, Richard, Hong, Paul, JP, Basten, Bas, Henri, Elio, Leila, Frank, Gabino.

Bedankt allemaal!

Lieve PaMa, het heeft even geduurd, maar jullie geduld is hiermee beloond.

Lieve Sylvia, Rosalinde en Julianne, aan het einde van de dag is het heerlijk thuiskomen bij jullie!

Curriculum Vitae

Harm Nijveen was born on January 13, 1971 in Hillegom (NL).

He received his MSc in chemistry at the University of Groningen in 1995. In that same year he started a PhD project at the University of Nijmegen.

In 1997 he temporarily left academia to join Applicare Medical Imaging B.V., which was acquired by General Electric in 1999. There he worked on software for radiology as software quality engineer and team leader of the software-testing group, and from 2000 as software engineer.

In 2001 Harm joined Dalicon B.V. as database manager of an indexing service for scientific literature. In 2002 he moved to the AMC hospital in Amsterdam to work as application engineer, maintaining the hospital's email and calendaring services, and various centrally managed medical software systems.

In 2004 Harm received his B ICT from the University of Applied Sciences Utrecht.

Since 2005 he is scientific programmer in the chair group of Bioinformatics in the Plant Sciences Group of Wageningen University. His responsibilities include teaching, coordinating the chair's education, research support, maintaining the computing facilities and software services.

In September Harm joined the Wageningen Seed Lab for two days per week as postdoc.

Publications

Alako BT, Rainey D, Nijveen H, Leunissen JA. 2006. TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure. *Nucleic Acids Res* **34**(Web Server issue): W104-109.

Ceccherini I, Hofstra RM, Luo Y, Stulp RP, Barone V, Stelwagen T, Bocciardi R, Nijveen H, Bolino A, Seri M. 1994. DNA polymorphisms and conditions for SSCP analysis of the 20 exons of the ret proto-oncogene. *Oncogene* **9**(10): 3025-3029.

Doerwald L, Nijveen H, Civil A, van Genesen ST, Lubsen NH. 2001. Regulatory elements in the rat betaB2-crystallin promoter. *Exp Eye Res* **73**(5): 703-710.

Gavai AK, Tikunov Y, Ursem R, Bovy A, van Eeuwijk F, Nijveen H, Lucas PJ, Leunissen JA. 2009. Constraint-based probabilistic learning of metabolic pathways from tomato volatiles. *Metabolomics* **5**(4): 419-428.

Hassan C, Kester MG, de Ru AH, Hombrink P, Drijfhout JW, Nijveen H, Leunissen JA, Heemskerk MH, Falkenburg JH, van Veelen PA. 2013. The Human Leukocyte Antigen-presented Ligandome of B Lymphocytes. *Mol Cell Proteomics* **12**(7): 1829-1843.

Hombrink P, Hassan C, Kester MG, de Ru AH, van Bergen CA, Nijveen H, Drijfhout JW, Falkenburg JH, Heemskerk MH, van Veelen PA. 2013. Discovery of T cell epitopes implementing HLA-peptidomics into a reverse immunology approach. *J Immunol* **190**(8): 3869-3877.

Kertesz-Farkas A, Dhir S, Sonego P, Pacurar M, Netoteia S, Nijveen H, Kuzniar A, Leunissen JA, Kocsor A, Pongor S. 2008. Benchmarking protein classification algorithms via supervised cross-validation. *J Biochem Biophys Methods* **70**(6): 1215-1223.

Kuzniar A, Dhir S, Nijveen H, Pongor S, Leunissen JA. 2010. Multi-netclust: an efficient tool for finding connected clusters in multi-parametric networks. *Bioinformatics* **26**(19): 2482-2483.

Kuzniar A, Lin K, He Y, Nijveen H, Pongor S, Leunissen JA. 2009. ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res* **37**(Web Server issue): W428-434.

Luo H, Lin K, David A, Nijveen H, Leunissen JA. 2012. ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Res* **40**(Database issue): D394-399.

Luo H, Nijveen H. 2013. Understanding and identifying amino acid repeats. *Briefings in Bioinformatics* doi: 10.1093/bib/bbt003.

Matus-Garcia M, Nijveen H, van Passel MW. 2012. Promoter propagation in prokaryotes. *Nucleic Acids Res* **40**(20): 10032-10040.

Nijveen H, Kester MG, Hassan C, Viars A, de Ru AH, de Jager M, Falkenburg JH, Leunissen JA, van Veelen PA. 2011. HSPVdb--the Human Short Peptide Variation Database for improved mass spectrometry-based detection of polymorphic HLA-ligands. *Immunogenetics* **63**(3): 143-153.

Nijveen H, Matus-Garcia M, van Passel MWJ. 2012. Promoter reuse in prokaryotes. *Mobile Genetic Elements* **2**(6): 6-8.

Nijveen H, van Kaauwen M, Esselink DG, Hoegen B, Vosman B. 2013. QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Res* **41**(Web Server issue): W587-590.

Stelwagen T, Hofstra RM, Stulp RP, Nijveen H, Romeo G, Landsvater RM, Lips CJ, Buys CH. 1994. Mutations in the RET proto-oncogene in sporadic medullary thyroid carcinomas. *Cancer Genetics and Cytogenetics* **77**(2): 175-175.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**(Web Server issue): W71-74.

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: Harm Nijveen
Date: 5 December 2013
Group: Bioinformatics, Wageningen University & Research Centre

1) Start-up phase	date
► **First presentation of your project**	
Network visualisation	Oct 22, 2009
► **Writing or rewriting a project proposal**	
Visualization of complex data	Dec 2, 2009
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*4.5 credits**

2) Scientific Exposure	date
► **EPS PhD student days**	
NBIC event: BioAssist programmers, social event Dalfsen	Sep 17, 2010
EPS PhD Student Day, Leiden University	Nov 29, 2013
► **EPS theme symposia**	
Theme 4 Symposium 'Genome Biology', Utrecht University	Dec 13, 2013
► **NWO Lunteren days and other National Platforms**	
NBIC conference 2009, Lunteren (NL)	Mar 17-18, 2009
NBIC conference 2010, Lunteren (NL)	Mar 29-30, 2010
NCSB Conference 2010, Lunteren (NL)	Oct 20-21, 2010
NBIC conference 2011, Lunteren (NL)	Apr 19-20, 2011
NCSB Conference 2011, Lunteren (NL)	Oct 31-Nov 01, 2011
NBIC conference 2012, Lunteren (NL)	Apr 24-25, 2012
NCSB Conference 2012, Lunteren (NL)	Nov 01-02, 2012
NBIC conference 2013, Lunteren (NL)	Apr 16-17, 2013
► **Seminars (series), workshops and symposia**	
EMBRACE Workshop: "Modern computer tools for the biosciences"	Mar 22, 2007
Oracle Data Mining and Search Seminar with Ron Hardman	May 29, 2007
Joint EBI / OMII-UK Workshop	Oct 08, 2007
Ask Tom Live: A 2-day Oracle Seminar with Tom Kyte	Jan 26, 2009
WEES seminar: Tal Dagan	Oct 22, 2009
WEES seminar: Michiel Vos	Mar 18, 2010
WEES seminar: Bas Haring	Sep 16, 2010
WEES seminar: Fiona Jordan	Nov 18, 2010
Wageningen Centre for Systems biology kick-off meeting	Apr 19, 2012
WEES seminar: Patrick Forterre	Oct 18, 2012
EPS Seminar Detlef Weigel	Feb 27, 2013
WEES seminar: Evolution in the laboratory (with Prof. Richard Lenski)	Mar 14, 2013
Applications of Illumina HiSeq: a new evolution in agriculture analysis	Apr 22, 2013
WAY public lecture Frans van Waal	Jun 26, 2013
► **Seminar plus**	
► **International symposia and congresses**	
ISMB/ECCB 2007 Vienna, Austria	July 21-25, 2007
Benelux Bioinformatics Conference, Maastricht, NL	Dec 15-16, 2008
2011 International Symposium on Integrative Bioinformatics, Wageningen, NL	Mar 21-23, 2011
► **Presentations**	
ISMB/ECCB 2007 Vienna (poster Primer3Plus)	Jun 21, 2007
BioRange meeting (Application show case)	Mar 05, 2008
Benelux Bioinformatics Conference (talk)	Dec 15, 2008
EMBO Workshop on Visualizing Biological Data VIZBI (poster)	Mar 03, 2010
NCSB 2010 symposium (talk)	Oct 20, 2010
NBIC BioAssist programmers meeting (talk)	Dec 17, 2010
NBIC conference 2011 (poster visualising complex networks)	Apr 19, 2011
NBIC conference 2012 (poster: LineUp Cytoscape plugin)	Apr 24, 2012
► **IAB interview**	
► **Excursions**	
Subtotal Scientific Exposure	*21.0 credits**

3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
Grid Tutorial (SARA/Surf)	Aug 25, 2007
Scientific Software as an Asset (NBIC)	Nov 04, 2008
CUDA programming TU Delft	Aug 04, 2009
EMBO Course 'Visualizing Biological Data VIZBI'	Mar 03-05, 2010
NCSB Systems Biology Tutorial: Basic Modelling	Jun 24, 2011
EPS course: Bioinformatics - a user's approach (course organiser)	2012, 2013 (2x)
EPS/NBIC course: The Power of RNA-seq (course organiser)	2013 (2x)
► **Journal club**	
WUR Bioinformatics literature discussions	2007-2012
► **Individual research training**	
Subtotal In-Depth Studies	*5.1 credits**

4) Personal development	date
► **Skill training courses (highly recommended)**	
Management voor aio's en oio's KU Nijmegen	Feb 02, 1997
Workshop "Valorisation: usability and exploitation" (NBIC conference)	Mar 05, 2008
Workshop "Work smart and save time" - Gerald Essers (NBIC conference)	Mar 29, 2010
Workshop "Open Access" by Phil Bourne and Jan Velterop (NBIC conference)	Apr 19, 2011
► **Organisation of PhD students day, course or conference**	
EPS course: Bioinformatics - a user's approach (course organiser)	Aug 27-31, 2012
EPS course: Bioinformatics - a user's approach (course organiser)	Mar 04-08, 2013
EPS course: Bioinformatics - a user's approach (course organiser)	Aug 26-30, 2013
MSc course: Advanced Bioinformatics 6 ECTs (course organiser) 20 days	2010
MSc course: Advanced Bioinformatics 6 ECTs (course organiser) 20 days	2011
MSc course: Advanced Bioinformatics 6 ECTs (course organiser) 20 days	2012
MSc course: Advanced Bioinformatics 6 ECTs (course organiser) 20 days	2013
EPS/NBIC course: The Power of RNA-seq (course organiser)	Jun 05-07, 2013
EPS/NBIC course: The Power of RNA-seq (course organiser)	Dec 16-18, 2013
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*4.8 credits**

TOTAL NUMBER OF CREDIT POINTS*	**35.4**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by the Netherlands Consortium for Systems Biology, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research, and by the Netherlands Bioinformatics Centre.