

## University of Groningen

### Haplotype resolved genomes

Porubský, David

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2017

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Porubský, D. (2017). *Haplotype resolved genomes: Computational challenges and applications*. University of Groningen.

#### **Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

#### **Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

# **Haplotype resolved genomes**

**Computational challenges and applications**

**David Porubský**

Cover design by David Porubsky

About the cover:

Most genomes including humans are diploid, with two homologous copies of each chromosome, one inherited from the mother and one from the father. These two homologous copies harbor a distinct set of genetic variants (single nucleotide variants – SNVs, inversions etc.). Collection of such variants along a single homologous chromosome is called a haplotype. Haplotype-resolved genomes are important in many areas of personalized medicine, population genetics as well as clinical research. In this thesis the power of strand sequencing (Strand-seq) is presented. Strand-seq is a single cell sequencing technique that can distinguish parental homologous base on the template strand inheritance in a single cell. This allows direct visualization and mapping of balanced rearrangements like inversions as well as the global phasing of genetic variants along the whole length of all chromosomes. Cover art was inspired by a popular arcade game from 80's called 'PAC-MAN'. We used the idea of this game to depict the main principle of haplotyping diploid genomes. Maternal and paternal homologue of Chromosome 1 are represented by Ms. and Mr. Hapman, respectively, both starting their haplotyping journey from opposing corners of the maze. Their goal is to find the way through the maze while connecting homologue specific SNVs and thus completing unique haplotypes.

Legal statement:

PAC-MAN™ copyrights and trademarks are proprietary to by BANDAI NAMCO Entertainment Inc. in TOKYO, 108-0014. Any use or reproduction of this cover for commercial purposes is strictly prohibited.

The printing of this thesis was financially supported by:  
University of Groningen and University Medical Center Groningen

ISBN (print version): 978-90-367-9637-8

ISBN (digital version): 978-90-367-9636-1

Copyright © 2016 by David Porubsky

No parts of this book may be reproduced or transmitted in any form or by any means without a permission of the author. All rights reserved.



university of  
 groningen

# Haplotype resolved genomes

Computational challenges and applications

PhD thesis

to obtain the degree of PhD at the  
 University of Groningen  
 on the authority of the  
 Rector Magnificus Prof. E. Sterken  
 and in accordance with  
 the decision by the College of Deans.

This thesis will be defended in public on

Monday 27 March 2017 at 16.15 hours

by

David Porubský

born on 1 February 1985  
 in Trenčín, Slovakia

**Supervisor**

Prof. P. M. Lansdorp

**Co-supervisors**

Dr. V. Guryev

Dr. M. R. Bevova

**Assessment committee**

Prof. E. Cuppen

Prof. M. A. T. M. van Vugt

Prof. J. Korbel

**Paranymphs**

Magda Grudniewska

Seka Simone Lazare

# Preface

The work described in this thesis is focused on the challenges and the promises of completely phased diploid genomes. Here I describe attractive solutions to this complex problem using single cell strand sequencing (Strand-seq). My first aim was to develop a robust computational framework for the assembly of accurate and global haplotypes from single cell Strand-seq libraries. My next objective was to validate my phasing approach in comparison to gold-standard haplotypes obtained from the HapMap project as well as to demonstrate the applicability of Strand-seq phasing. Lastly, I investigated integrative phasing, a combination of Strand-seq with parallel sequencing technologies, to reduce the costs and increase the completeness of haplotype resolved genomes.

A general introduction is provided in **Chapter 1** which represents a comprehensive summary of current and past haplotyping techniques. I highlight the importance of haplotype information in basic and clinical research. I emphasize emerging technologies like single cell sequencing, linked-read sequencing as well as long-read technologies and outline their capabilities to provide haplotype-resolved genomes.

Strand-seq is a unique single cell sequencing technique with a wide range of applications. However, the computational tools required to interpret Strand-seq data are still in their infancy. In **Chapter 2** I describe the development of novel bioinformatics pipelines capable of handling subtle nuances of Strand-seq data. Specifically, I describe and discuss the development of tools for the mapping of breakpoints and haplotype phasing using Strand-seq data.

The validation of these tools is presented in **Chapter 3**. For this purpose we have chosen a well-known family trio from the HapMap project as well as other independent data sources like PacBio RNA-seq and other single cells phasing method. This work revealed high accuracy of Strand-seq phasing which was further highlighted in comparison to *de novo* assembly based phasing.

In **Chapter 4** I describe the application of genome-wide haplotyping using Strand-seq to map meiotic recombination events in a family trio as well as haplotype differences at the single cell level. Furthermore, I demonstrate the phasing of larger

genetic variants like deletions, duplications and inversions.

In order to reduce the cost and labor need to phase the genome of a single individual I have explored the possibilities to integrate global phasing provided by Strand-seq with other sequencing technologies like PacBio or Illumina. In **Chapter 5** I present such integrative phasing approach combining global Strand-seq haplotypes with local haplotypes embedded in sequenced DNA fragments.

**Chapter 6** provides a summary of the results presented in this thesis, along with general discussion on major implications of these results to the scientific community. I further discuss directions and recommendations to achieve *de novo* assembled and phased individual genomes. Future perspectives, and work ahead is discussed as well.

# Contents

<b>Chapter 1 </b>		<b>10</b>
General Introduction: Genetic variation of diploid genomes unveiled [Manuscript in preparation]		
<b>Chapter 2 </b>		<b>24</b>
Single cell analysis toolkit: BreakPointR, StrandPhase and StrandPhaseR [Manuscript in preparation]		
<b>Chapter 3 </b>		<b>42</b>
Direct chromosome-length haplotyping by single cell sequencing (PART 1) [Genome Research 2016; 10.1101/gr.209841.116]		
<b>Chapter 4 </b>		<b>62</b>
Direct chromosome-length haplotyping by single cell sequencing (PART 2) [Genome Research 2016; 10.1101/gr.209841.116]		
<b>Chapter 5 </b>		<b>84</b>
Integrative experimental and read-based phasing of a single individual [Manuscript in preparation]		
<b>Chapter 6 </b>		<b>108</b>
Discussion and future perspectives		
<b>Appendices</b>		
	Bibliography	<b>117</b>
	Dutch summary	<b>124</b>
	Glossary	<b>126</b>
	List of abbreviations	<b>129</b>
	Acknowledgements	<b>130</b>
	Curriculum Vitae	<b>134</b>
	List of publications	<b>136</b>

HOGGIE

INSERT A COIN  
&  
PRESS START

