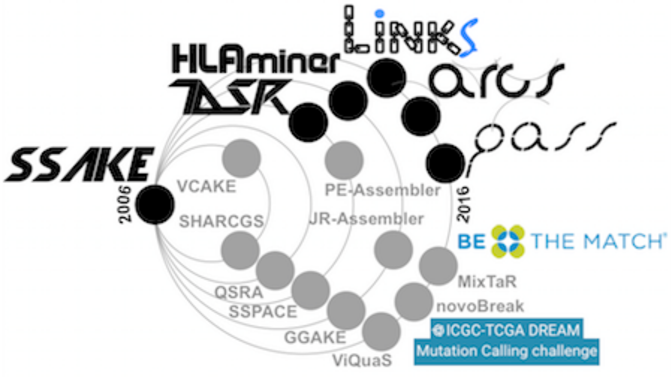
# René L. Warren

## Personal Statement

René is a researcher with extensive experience in computational biology, a prolific scientific research author and regular reviewer for peer reviewed genomics and bioinformatics journals. In his career, he has been at the forefront of innovation, and pioneered a number of bioinformatics “firsts”. Among those, SSAKE, the first algorithm for *de novo* genome assembly with short DNA sequences [1]. Algorithms of SSAKE are the core of many genomics applications and their design continues to inspire new-generation technologies. Applications of the software extend beyond genome assembly; The innovative technology was applied to profiling T-cell metagenomes [12], targeted *de novo* genome sequence assembly (TASR) [5], HLA typing (HLAminer [16] and NMDP Be The Match®), genome scaffolding with long reads (LINKS) [3], proteome assembly (PASS) and was key to the discovery of *Fusobacterium* in colon cancer [9], a finding designated as one of the top 10 medical breakthroughs of 2011 by Time magazine. The bioinformatics technologies he developed are actively maintained and supported. A complete list of his published work is available from https://scholar.google.ca/citations?user=nLkZYtcAAAAJ&hl=en

## Contributions to Science

### I. Development of the first *de novo* genome assembler for next-generation DNA sequences



Sample applications derived from *SSAKE*

In September 2006, René designed and developed the first algorithm for very short, 23nt, Solexa reads shortly before Illumina acquired the company. Next-generation sequencing was nascent, not yet widespread at the time and its applicability for *de novo* genome assembly was questioned due to the short read length. René saw an opportunity to enable scientists, providing the means to assemble viral genomes *de novo* and kilobase-long sequences that would facilitate the characterization of DNA and RNA samples. The innovative algorithm, called SSAKE, was initially built on the premise that massively parallel sequencing would oversample good sequences, easily distinguishing them from errors (noise) in the data thus facilitating the assembly process [1,2]. The software quickly spurred the development of similar technologies based on the open-source SSAKE code, including VCAKE and SHARGCS that followed the same algorithmic paradigm one year later (Figure 1). Since inception over 10 years ago, the software has matured and improved to handle base errors in reads and includes paired-end read logic for assembly and scaffolding. The SSAKE scaffolder was later modified by the company BaseClear in the Netherlands and gave rise to the popular stand-alone SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) scaffolder. In 2015, René extended the scaffolder to handle error-rich long reads from emerging Oxford Nanopore Technologies and established Pacific Biosciences DNA sequence platforms, spearheading the development of LINKS, a scalable, alignment-free genome assembly scaffolder [3,4]. The SSAKE technology has had significant impacts in biology and is often cited (over 475 google scholar citations). Some of the new *de novo* short-read assemblers, including the JR-assembler and PE-assembler follow on the same algorithmic designs. SSAKE is also an integral component and the central engine behind a wide array of tools such as quasispecies assembler ViQuaS, tandem repeat detection MixTaR and HLAminer, the first and widely-used Human Leukocyte Antigen (HLA) prediction software for sequence shotgun data written by René (see below). It is also the *de novo* software of choice currently integrated in the HLA prediction software pipeline built by the U.S. National Marrow Donor Program (Be The Match®) and, considering the central role of HLA in adaptive immunity, has far-reaching health applications. Further, its use was instrumental in the assembly of specific contig targets upon which primer-probe sets were designed to identify, for the first time, bacterium *Fusobacterium nucleatum* in colorectal cancer (CRC) tumor samples (see below). Further, SSAKE is the assembly engine in the top-performing cancer genomic structural variant predictor pipeline software novoBreak as assessed by The ICGC-TCGA DREAM Genomic Mutation Calling Challenge (Chong *et al.*, 2016)

1. Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*. 23:500-501. PMID17158514.

2. http://www.bcgsc.ca/platform/bioinfo/software/ssake

3. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*. 4:35. PMID26244089.

4. http://www.bcgsc.ca/platform/bioinfo/software/links

**II. Development of the first targeted de novo assembler for next-generation DNA sequences**

As a natural extension of SSAKE, René designed and developed the first bioinformatics tool for Targeted Assembly of Sequence Reads, TASR, which was published in 2011 [5,6]. For applications where a whole genome or transcriptome assembly is not necessary or even feasible, it is judicious to assemble only sequences of interest. For instance, these could be transcripts, gene fusions, virus integration sites and sequences capturing variants of interest that need to be validated in new samples. The TASR algorithm is designed to quickly interrogate shotgun DNA sequence data from large cohorts quickly, using a hypothesis-driven approach. The innovation in TASR is the deconstruction of sequences of interest – targets – into k-mers that are hashed and used to recruit whole genome or transcriptome sequence reads when a 5'-end matching k-mer is found, effectively binning subset of reads for target-assisted or *de novo* assembly. This approach is precursor to the methodology proposed in our present application. TASR is the engine behind René's HLA prediction software HLAminer (below) and was used to profile the expression of the previously uncharacterized long non-coding RNA (lncRNA) *EVADR* in cancers of glandular origins, using ~7500 patient sample RNA-seq data from The Cancer Genome Atlas (TCGA) [7].

5. Warren RL, Holt RA. 2011. Targeted assembly of short sequence reads. *PLoS ONE* 6(5):e19816. PMID21589938.

6. http://www.bcgsc.ca/platform/bioinfo/software/tasr

7. Gibb EA, Warren RL, Wilson GW, Brown SD, Robertson G, Morin GB, Holt RA. 2015. Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. *Genome Med*. 7:22. PMID25821520.

**III. Discovery, for the first time, of the association between *F. nucleatum* and colorectal cancer**

With the etiology of a growing number of cancers being linked to microbial exposure such as HPV and *H. pylori* in many cervical and gastric cancers, and the recent discoveries of complex microbial communities living along the gastro-intestinal (GI) tract, Robert Holt, Richard Moore and René Warren looked at possible links between microbes and CRC. Providing bioinformatics expertise on the project, René designed and implemented a pathogen detection pipeline, which he used to assess the detection limit of DNA sequencing for the identification of probable infectious agents [8]. In this initial and controlled experiment, spiked-in RNA viruses in experimental CRC samples were detected in concentrations as low as 0.1 parts per million. Subsequently, using a dozen tumor and matched adjacent normal samples from that many CRC patients, René discovered, for the first time, statistically significant elevated sequence signatures of *F. nucleatum* in the tumor transcriptomes of CRC patients, using his infectious agent detection pipeline [9]. He subsequently assembled *de novo* all *F. nucleatum* reads in each patients and identified *F. nucleatum* specific genes that were ultimately utilized to design primer-probe sets for qPCR, which were in turn used to quickly screen a larger cohort of 99 CRC patients and became the basis of a patent to detect *Fusobacterium* in GI samples [10,11]. The *F. nucleatum* discovery in CRC was independently made by the group of Matthew Meyerson at the Broad

Institute of MIT and Harvard (Kostic *et al.*, 2011) and, in 2011, it was designated as one of the top 10 medical breakthroughs of the year by Time magazine. The human health implication of this discovery is paramount and has spearheaded international collaborations with scientists at the BC Cancer Agency. It led to a diagnostic test for CRC [10], identified a possible etiological agent of CRC and, in the future, may provide actionable targets that could be exploited to prevent infection, inflammation and eventual development of CRC.

8. Moore RA*, Warren RL*, Freeman JD, Gustavsen JA, Chénard C, et al. 2011. The Sensitivity of Massively Parallel Sequencing for Detecting Candidate Infectious Agents Associated with Human Tissue. *PLoS ONE* 6(5):e19838. PMID21603639. *authors contributed equally

9. Castellarin M*, Warren RL*, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA. 2012. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*. 22:299-306. PMID22009989. *authors contributed equally

10. E Allen-Vercoe, R Holt, R Moore, R Warren - US Patent App. 13/877,421, 2011. Detection of fusobacterium in a gastrointestinal sample to diagnose gastrointestinal cancer.

11. Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, Allen-Vercoe E, Holt RA. 2013. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*. 1:16. PMID24450771.

**IV. Development of the first bioinformatics pipeline to profile T cell receptor sequences**

T cells mature in the thymus where the genome undergoes rearrangement of the variable, joining and diversity (VDJ) T Cell Receptor (TCR) gene segments to help create the diversity T cells need to recognize combinations of HLA-presented epitopes at the cell surface and mount an appropriate immune response. It is estimated that there exist $10^{15}$ theoretical receptor for the TCR beta chain alone. We sought to profile the TCR metatranscriptome in healthy individuals and measure precisely the total number of T cell clonotypes observed based on their expressed receptor, captured using custom amplification of TCR transcripts. René designed and developed the bioinformatics approaches needed to support the next-generation sequencing spectratyping program. Initially the sequence profiling utilized a derivation of the SSAKE algorithm, immunoSSAKE [12,13], to assemble very short reads (36-50nt) at the VDJ junction, but soon parted from this methodology as the read lengths increased and was replaced by a micro-assembler of his design for overlapping paired-end long reads from the same amplicon [14]. The laboratory methods for TCR amplicon amplification and sequencing and the bioinformatics pipeline for characterizing TCR and B cell receptor sequences are now both in production at the BC Cancer Agency, Genome Sciences Centre and was used to identify lobular breast tumor-infiltrating T cells [15]. As read length increase it will become easier to annotate immune receptor sequences, and reporting on the composition of T cells has tremendous value for clinicians. For instance, the ability to characterize the T cell repertoire at nucleotide resolution is especially helpful to accurately monitor the immune repertoire of patients post organ transplantation.

12. Warren RL, Nelson BH, Holt RA. 2009. Profiling model T cell metagenomes with short reads. *Bioinformatics*. 25:458-464. PMID19136549.

13. Freeman JD*, Warren RL*, Webb JR, Nelson BH, Holt RA. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res*. 19:1817-1824. PMID19541912. *authors contributed equally

14. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res*. 21:790-797. PMID21349924.

15. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K,

Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S. 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*. 461:809-813. PMID19812674.

**V. Development of HLAminer, the first software for HLA prediction from shotgun sequence data**

The Human Leukocyte Antigen (HLA) is a gene locus that encodes cell-surface receptors that present epitopes to T cells, and is the foundation of immune system regulation in humans. The HLA complex enables the immune system to distinguish between a person's normal healthy cells, and cells that are infected (for example with a virus), mutated (for example in cancer) or are derived from some other individual (for example, upon organ transplant). Due to the necessity of recognizing diverse molecular signatures, there is extreme diversity in the genes that encode HLA proteins, and determining the specific gene variants of an individual is of considerable importance in clinical medicine and in research. For instance, knowing HLA genes (types) and matching donor and recipients with the same HLA is key to successful organ grafts. Further, HLA genes are informative and often the determinants of disease. Traditionally, HLA typing was done by PCR amplification, but often offers low allele-level resolution. The drastic decrease in sequencing costs we witnessed over the past decade is facilitating the uptake of DNA sequencing technologies in the clinic, but the bioinformatics analysis of whole genome and transcriptome shotgun data set still represents a significant barrier to entry and is out of reach to many researchers and clinicians. As more shotgun data sets become available, mining them to predict HLA types is value-added since those datasets can be mined retrospectively. René designed and implemented the first HLA prediction software for next-generation shotgun sequence data. The flexible pipeline enables HLA predictions from direct read alignments and, for best results, uses TASR as its *de novo* HLA sequence assembly engine upon which predictions are based. The software was used to robustly predict the HLA genes of the colorectal cancer patient cohort described above and in over 7000 TCGA RNA-seq samples (of which 515 from six tumor sites were published) [16,17,18]. In 2016 alone, the HLAminer software has had over 17,880 downloads world-wide.

16. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA. 2012. Derivation of HLA types from shotgun sequence datasets. *Genome Med*. 4:95. PMID23228053.

17. http://www.bcgsc.ca/platform/bioinfo/software/hlaminer

18. Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, Holt RA. 2014. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*. 24:743-50. PMID24782321.