Development of integrative bioinformatics resources for the analysis of viral next generation sequencing (NGS) data and human papillomaviruses (HPVs)

THESIS

Submitted for the degree of

DOCTOR OF PHILOSOPHY

TO

JAWAHARLAL NEHRU UNIVERSITY

NEW DELHI, INDIA

2020

AMIT KUMAR GUPTA

INSTITUTE OF MICROBIAL TECHNOLOGY (CSIR-IMTech)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH
SECTOR 39A, CHANDIGARH-160036, INDIA

NOVEMBER 2020

सीएसआईआर – इमटैक
CSIR-IMTECH

सीएसआईआर – सूक्ष्मजीव प्रौद्योगिकी संस्थान
सैक्टर 39 – ए, चण्डीगढ़ – 160 036 (भारत)
CSIR-INSTITUTE OF MICROBIAL TECHNOLOGY
(A CONSTITUENT ESTABLISHMENT OF CSIR)
Sector 39-A, Chandigarh-160 036 (INDIA)

CERTIFICATE

The research work embodied in this thesis entitled "***Development of integrative bioinformatics resources for the analysis of viral next generation sequencing (NGS) data and human papillomaviruses (HPVs)***" has been carried out by Mr. Amit Kumar Gupta under the guidance of Dr. Manoj Kumar, at the Institute of Microbial Technology (CSIR-IMTech), Sector 39A, Chandigarh. This thesis is an original work and has not been submitted, in part or as a whole, for a degree at this or any other university. Nor does it contain, to the best of our knowledge and belief, any material published or written by any other person, except as acknowledged in the text.

Manoj Kumar

Dr. Manoj Kumar
(Supervisor)

डॉ. मनोज कुमार/Dr. Manoj Kumar
प्रधान वैज्ञानिक/Principal Scientist
सीएसआईआर–सूक्ष्मजीव प्रौद्योगिकी संस्थान
CSIR-Institute of Microbial Technology
सैक्टर/Sector 39-A, चण्डीगढ़/Chandigarh-160036

Gupta

Amit Kumar Gupta
(Research scholar)

दूरभाष 0172-6665201 } Reception
0172-6665202 }
0172-6665..... (Direct)

फैक्स : { 0091-172-2690632 (COA)
Fax : { 0091-172-2690585 (Director)
0091-172-2690056 (Purchase)
एस. टी. डी. कोड : STD CODE : 172

तार : Gram : IMTECH
E.mail : system@imtech.res.in
Web : http://imtech.res.in

ACKNOWLEDGEMENTS

I would like to express my earnest gratitude to my advisor **Dr. Manoj Kumar** for letting me join his lab as a Ph.D. scholar in 2014 at Bioinformatics Centre (BIC), CSIR-IMTECH to pursue my research work. Sir, I am very thankful to you for all the guidance, constructive discussions, motivation, support and fostering throughout my tenure and work. I have learned and admired many things from you, scientific ethics, patience, positivity, optimism are just a few names. It was a great learning experience from you. Thank you so much Sir.

I would also like to give my sincere thanks to **Dr. G.P.S. Raghava** for invaluable guidance, teaching and inspiring me towards scientific goals. It was always a pleasure to work and learn from you. I will also take this opportunity to acknowledge him for constructing and providing excellent computational infrastructure facility at BIC. I am also thankful to **Dr. Balvinder Singh**, and **Dr. Anshu Bhardwaj** for the encouragement, teaching and positive advices during PhD work.

I also like to thank IMTECH IT team, **Harvinder Jassal Sir**, Harminder Sir, Sandeep Sir, Paramjit Sir, Chander Shekhar Sir and Amit for their constant support during my work. They are very efficient and dedicated towards maintaining the excellent computational and networking facility in IMTECH. I enjoyed healthy discussions especially with Harvinder, Sandeep, and Paramjit sir on public issues, news and IT topics.

I also take this opportunity to thank and acknowledge current and former directors of CSIR-IMTECH, **Dr. Sanjeev Khosla**, **Dr. Anil Koul**, and **Dr. Girish Sahni** for providing the excellent infrastructure and working environment at the institute. I also like to express my sincere thanks to Dr. Pradeep Chakraborti, Dr. Karthikeyan Subramanian, Dr. Charu Sharma, and Dr. Balwinder Singh for timely and effortlessly coordinating the JNU-IMTECH Ph.D. program. I am also thankful to the Ph.D. student office (Ms. Shashi Batra Ma'am, Mr. Jankey Sir) and the PTM division (Mr. Kailash T. Bhamare, Mr. Pradeep Kumar Patel) for helping us in official works and publications.

Further, I like to acknowledge my committee (including Research Advisory Committee) members (Dr. G.P.S. Raghava, Dr. Srikrishna Subramanian, Dr. Balvinder Singh, Dr. Anshu Bhardwaj, Dr. Manoj Raje, Dr. Krishan Gopal Thakur, Dr. Prabhu B. Patil), who gave

constructive suggestions and feedbacks towards improvement of my work during presentations and meetings. I am also grateful to Dr. S. Mayilraj for collaborative work. I would also like to thank Dr. Ashish for his support and friendly talks.

I would like to dedicate my thanks to all my current and previous lab members Dr. Abid, Dr. Nishant, Dr. Akanksha, Dr. Isha, Dr. Shoaib, Dr. Showkat, Dr. Karambir, Anamika, Gazaldeep, Shivangi, Amber, Adhip, Sakshi Kamboj, Dr. Vinay, Dr. Shivalika, Dr. Archit, Dr. Kirti, Dr. Himani, Vandna, Pallawi, Shubham, Manmeet, Barkha, Sakshi, Nishant, Chanchal, Shalu, and Aman for making this journey memorable and become so helpful to me. I would especially like to mention and thank Dr. Shoaib, Dr. Showkat, Anamika, and Dr. Karambir for all the support, encouragement, fun trips and cherished moments. Further, I also like to thank all the seniors, juniors and friends from the department, Dr. Bharat, Dr. Jagat, Dr. Ravi, Dr. Arun, Dr. Shailesh, Dr. Sudheer, Dr. Kumardeep, Dr. Rahul, Dr. Sandeep Dhanda, Dr. Deepak, Dr. Ankur, Dr. Gandharva, Dr. Deepika, Dr. Sangita, Dr. Sandeep Singh, Dr. Piyush, Pragya, Dr. Salman, Rajesh, Dr. Sherry, Vinod, Dr. Manika, Rakesh, Tina, Harpreet, Ayesha, and Meenu for time to time help and support. Also, thanks to my collaborators Dr. Sarabjeet, and Dr. Narender.

I want to express my heartfelt thankfulness to my companions Dr. Amar, Shekhar, Nitish, Pravinkumar, Dr. Piyush, Dr. Manoj for fond memories, bike trips and fun time at IMTECH from the very first day to till date. Special thanks to my roommate, Dr. Amar for all the great time and support. I also like to extend my thanks to all (SB14) batchmates, Dr. Amar, Shekhar, Nitish, Pravinkumar, Dr. Piyush, Dr. Manoj, Dr. Anil, Pragya, Manjula, Krishna, Dr. Prabhat, Dr. Ajit, Dr. Krushna, Pradeep, Prashant, Anand, Dr. Yahshpal, Dr. Virender, Avantika, Dr. Sonal, Dr. Surbhi, and Ramita. Further thanks to Rahul Dilawari, Nitesh, Dr. Shiv, Kautilya, Dr. Mangesh, Dr. Arun, Rahul Mishra, Manoj, Umesh, Navneet, Manjunath, Dr. Shikha, Dr. Aasawari, Dr. Gaurav Sharma, Dr. Pankaj Sharma, Dr. Pradeep Rai, Dr. Sristy Shikha, Dr. Zeeshan, Dr. Aurobind, Mr. Vikas, Dr. Sajid Nadeem, Dr. Abhishek Trivedi, Dr. Himanshu Malhotra, Dr. Rehan, Dr. Surendra Vikram, Dr. Ghanshyam Yadav, Sachin, Jitender, B.N. Shukla, Navjot, Lucky, Yogita, Vaidhvi, Neha, Kanti, Hina, Poushali, Naushad, Gaurav Chaubey, Manish, Rahis, Sahil, other Seniors and Juniors (the list is endless) at IMETCH, for all the friendly and lively interactions and helping me out in one way or another.

I also want to acknowledge my volleyball team members and SSBMT participants, Dr. Pradeep Mishra, Dr. Jagpreet, Dr. Sandeep Dhanda, Dr. Shailendra, Dr. Abhijit, Dr. Rakesh, Mr. Asheesh Kumar Khare, Mr. Ashok Rana, Mr. Anil Kumar Sharma, Mr. Rakesh Kumar Dhiman, and Mr. Jaideep Mehta.

I sincerely acknowledge the **Department of Science and technology (DST)**, Govt. of India for providing me **INSPIRE fellowship** for financial assistance during the Ph.D. tenure. Council of Scientific & Industrial Research (CSIR) and Department of Biotechnology (DBT) is also acknowledged for financial support and other expenditures.

I also like to thank the staff of administration, finance, accounts, library, ESD, store and purchase for all the support and corresponding services. I would also like to extend my thankfulness to staff of hostel facility, mess and cafeteria, security for effortless work, and horticulture for maintaining lush green campus.

This acknowledgment stands incomplete without the mention of my wonderful Family. My special thanks to my Father **Sh. Jagdish Chandra Gupta** and Mother **Smt. Seema Gupta** for all the sacrifices they made, unconditional love, care and for supporting and teaching me in every step of my life. However, no words will be enough to thank them. They are the inspiration and power behind my every achievement. During this journey, I also met with my lifetime companion. My heartfelt thanks to my Wife, **Aishwarya Gupta**, who stood by my side during all the ups and downs. She gave me all the moral support, care and love. She is also very caring towards family. God also blessed us with a lot of happiness in the form of Daughter **Vedanshi** (My Angel). Her smile makes me forget all the worries and give us more reasons to love life. I would also like to thank my Sister **Samta** and brother-in-law for all the love and support. Special love to my two cute nieces- Ishita and Pari. I would also like to thank my family-in-law, cousins, and extended family (Mama, Mami, Masi, Masaji, Bhua, Fufaji). Thanks for everything. I would like to dedicate my entire work to the family, all teachers and friends.

At Last but not the least, my greatest regard to Almighty God for everything I have, family, friends, teachers, good health, morality and wisdom.

-Amit Kumar Gupta

ABBREVIATIONS

Abbreviations	Full Form
AAC	Amino acid composition
ACC	Accuracy
ANN	Artificial neural network
AVPs	Anti-viral peptides
BAM	Binary alignment map
BIN	Binary
BLAST	Basic Local Alignment Search Tool
BP	Biological process
BWT	Burrows-Wheeler transform
CaCx	Cervical carcinoma
CAS	CRISPR-associated proteins
CC	Cellular components
CD	Cell differentiation
CD4	Cluster of Differentiation 4
CD8	Cluster of Differentiation 8
CDS	Coding sequence
CESC	Cervical squamous cell carcinoma and Endocervical Adenocarcinoma
CIN	Cervical intraepithelial neoplasia
CNVs	Copy number variations
CRISPR	Clustered regularly interspaced short palindromic repeats
CSCC	Cervical squamous cell carcinoma
CSS	Cascading style sheets
CTL	Cytotoxic T-lymphocyte
dBg	de Bruijn graph
DEGs	Differential expressed genes
DENV	Dengue virus
DNA	Deoxyribonucleic acid
DPC	Di-peptide composition

Abbreviations	**Full Form**
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
EBV	Epstein-Barr virus
ENA	European Nucleotide Archive
EPC	Edge percolated component
ERVs	Endogenous retroviruses
EVEs	Endogenous viral elements
FN	False negative
FP	False positive
FPR	False positive rate
GDC	Genomic Data Commons
GFF	Gene feature file
GLOBOCAN	Global cancer observatory
GO	Gene ontology
GSEA	Gene set enrichment analysis
GVP	Global Virome Project
HBV	Hepatitis B virus
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus
HHV4	Human herpesvirus 4
HHV8	human herpesvirus 8
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HNSCC	Head neck squamous cell carcinoma
HP	Human disease phenotypes
HPVs	Human papillomaviruses
HR-HPVs	High-risk HPVs
HTLV-1	Human T-cell lymphotrophic virus type 1
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HVPC	Human Virome Protein Cluster Database

Abbreviations	**Full Form**
HyperM	Hyper methylation
HypoM	Hypo methylation
IARC	International Agency for Research on Cancer
ICC	Invasive cervical cancer
IDBA	Iterative De Bruijn graph Assembler
IEDB	Immune Epitope Database
IVA	Iterative Virus Assembler
JSON	JavaScript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KSHV	Kaposi`s sarcoma-associated herpesvirus
LAMP	Linux-Apache-MySQL-PHP
LCR	Long control region
LR-HPVs	Low-risk HPVs
MCC	Mathew's correlation coefficient
MF	Molecular functions
MHC	Major histocompatibility complex
miRNAs	MicroRNAs
MLTs	Machine learning techniques
MNC	Maximum neighbourhood component
MSigDB	Molecular Signatures Database
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Next generation sequencing
NIH	National Institutes of Health
NLM	National Library of Medicine
nt	Nucleotide
OLC	Overlap, Layout, Consensus
oncomirs	Oncogenic miRNAs
ONT	Oxford Nanopore Technologies
ORFs	Open reading frames
OS	Operating system

Abbreviations	**Full Form**
PAM	Protospacer Adjacent Motif
PaVE	Papillomavirus Episteme
PCR	Polymerase chain reaction
PERL	Practical Extraction and Report Language
PHP	Hypertext Preprocessor
PPI	Protein-protein interaction
QC	Quality control
Rb	Retinoblastoma
RNA	Ribonucleic acid
ROC	Receiver operating curve
SAM	Sequence alignment map
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
sgRNAs	Single guide RNAs
siRNAs	Small interfering RNAs
SRA	Sequence Read Archive
ssRNA	Single-stranded RNA
STRING	The Search Tool for the Retrieval of Interacting Genes
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
TFs	Transcription factors
TN	True negative
TP	True positive
TPR	True positive rate
ViPR	Virus Pathogen Database and Analysis Resource
VLPs	Virus like particles
WNV	West Nile virus
WWW	World Wide Web

Figure Legends

Chapter 1. Introduction

Chapter 2. Development of HPV mediated disease biomarker knowledgebase

Chapter 5. Benchmarking of de novo genome assemblers for the viral next generation sequencing (NGS) data

Chapter 6. Development of bioinformatics tool or pipeline for viral NGS data analysis

Table Legends

Chapter 1. Introduction

Chapter 2. Development of HPV mediated disease biomarker knowledgebase

Chapter 3. Systematic meta-analysis of human genes disrupted due to HPVs associated events

Chapter 4. Development of human papillomavirus (HPV) genomic and therapeutic resource

PUBLICATIONS

Thesis

Publications:

- ✓ **Amit Kumar Gupta** and Manoj Kumar*. (2015) HPVbase- a knowledgebase of viral integrations, methylation patterns and microRNAs aberrant expression: As potential biomarkers for Human papillomaviruses mediated carcinomas. **Sci. Rep.** 5, 12522. **(Impact Factor-5.578)**
- ✓ **Amit Kumar Gupta** and Manoj Kumar*. (2020) HPVomics: an integrated resource for the human papillomavirus epitome and therapeutics. **Genomics** 112 (6) 4853-4862. **(Impact Factor-6.205)**
- ✓ **Amit Kumar Gupta** and Manoj Kumar*. Multi-omics approach towards identification and analysis of therapeutic targets involved in HPV pathogenesis with special focus on carcinomas: Implication in drug repurposing (Under communication)
- ✓ **Amit Kumar Gupta** and Manoj Kumar*. Benchmarking of *de novo* genome assemblers on the viral next generation sequencing (NGS) data (Under communication)
- ✓ **Amit Kumar Gupta** and Manoj Kumar*. VIRpipe: an integrated pipeline for rapid virus identification and discovery from the clinical and environmental metagenomic samples (Under communication)

Posters:

- ✓ **Amit Kumar Gupta** and Manoj Kumar*. Landscape of HPV mediated events as potential biomarkers in diverse carcinomas. Proceedings of VIROCON 2015, XXIV National Conference of Indian Virological Society (IVS) at NEIGRIHMS, Shillong, India
- ✓ **Amit Kumar Gupta** and Manoj Kumar*. HPV integration associated genome-wide disruption –A functional and network analysis. Proceedings of NextGen Genomics, Biology and Bioinformatics and Technologies (NGBT) International Conference 2016 organized by SciGenom Research Foundation (SGRF) at Cochin, India

- ✓ **Amit Kumar Gupta** and Manoj Kumar*. Pilot study to evaluate the effect of different sequencing platforms and virus species on genome assembly quality. IMTechCon: An Industry- Academia meet, 2017, at CSIR-Institute of Microbial Technology, Chandigarh, India

Contributing works

Publications:

- ✓ Md Shoaib Khan[#], **Amit Kumar Gupta**[#] and Manoj Kumar*. (2015) ViralEpi v1.0: a high-throughput spectrum of viral epigenomic methylation profiles from diverse diseases. **Epigenomics**: 67-75. **(Impact Factor-4.044)**
- ✓ Akanksha Rajput, **Amit Kumar Gupta** and Manoj Kumar*. (2015) Prediction and Analysis of Quorum Sensing Peptides Based on Sequence Features. **PLoS ONE** 10(3): e0120066. **(Impact Factor-3.057)**
- ✓ Karambir Kaur, Himani Tandon, **Amit Kumar Gupta** and Manoj Kumar*. (2015) CrisprGE: a central hub of CRISPR/Cas-based genome editing. **Database (Oxford)**, bav055. **(Impact Factor-2.627)**
- ✓ **Amit Kumar Gupta**[#], Karambir Kaur[#], Akanksha Rajput[#], Sandeep Kumar Dhanda[#], Manika Sehgal[#], Md Shoaib Khan[#], Isha Monga, Showkat Ahmad Dar, Sandeep Singh, Gandharva Nagpal, Salman Sadullah Usmani, Anamika Thakur, Gazaldeep Kaur, Shivangi Sharma, Aman Bhardwaj, Abid Qureshi, Gajendra Pal Singh Raghava, and Manoj Kumar*. (2016) ZikaVR: An Integrated Zika Virus Resource for Genomics, Proteomics, Phylogenetic and Therapeutic Analysis**. Sci. Rep.** 6, 32713. **(Impact Factor-5.228)**
- ✓ Showkat Ahmad Dar, **Amit Kumar Gupta**, Anamika Thakur, and Manoj Kumar*. (2016) SMEpred workbench: A web server for predicting efficacy of chemically modified siRNAs. **RNA biology** 13 (11), 1144-1151. **(Impact Factor-4.076)**
- ✓ Karambir Kaur[#], **Amit Kumar Gupta**[#], Akanksha Rajput[#], and Manoj Kumar*. (2016) ge-CRISPR - An integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system. **Sci. Rep.** 6, 30870. **(Impact Factor-5.228)**
- ✓ Isha Monga, Abid Qureshi, Nishant Thakur, **Amit Kumar Gupta**, and Manoj Kumar*. (2017) ASPsiRNA: A Resource of ASP-siRNAs Having Therapeutic Potential for Human Genetic Disorders and Algorithm for Prediction of Their

Inhibitory Efficacy. **G3: Genes, Genomes, Genetics** 7 (9), 2931-2943. **(Impact Factor-2.742)**

- ✓ **Amit Kumar Gupta**, Archit Kumar, Akanksha Rajput, Karambir Kaur, Anamika Thakur, Showkat Ahmed Dar, Kirti Megha and Manoj Kumar*. (2020) NipahVR: multi-targeted putative therapeutics and epitome resource for the nipah virus. **Database**, Volume 2020, baz159. **(Impact Factor-2.593)**
- ✓ **Amit Kumar Gupta**[#], Md. Shoaib Khan[#], Shubham Choudhury[#], Adhip Mukhopadhyay[#], Sakshi[#], Amber Rastogi[#], Anamika Thakur, Pallawi Kumari, Manmeet Kaur, Shalu, Chanchal Saini, Vandna Sapehia, Barkha, Pradeep Kumar Patel, Kailash T. Bhamare and Manoj Kumar*. (2020) CoronaVR: A Computational Resource and Analysis of Epitopes and Therapeutics for Severe Acute Respiratory Syndrome Coronavirus-2. **Front Microbiol.** 2020; 11: 1858. **(Impact Factor-4.235)**
- ✓ Narender Kumar[#], **Amit Kumar Gupta**[#], Sarabjeet Kour Sudan[#], Deepika Pal, Vinay Randhawa, Girish Sahni, Shanmugam Mayilraj* and Manoj Kumar*. Abundance and diversity of phages, microbial taxa and antibiotic resistance genes in the sediments of the river Ganges through metagenomic approach (Under communication)
- ✓ Md Shoaib Khan, **Amit Kumar Gupta** and Manoj Kumar*. TcellEpi: Prediction of immunogenic and non-immunogenic CD8+ T-cell and CD4+ T-cell epitopes (Under communication)
- ✓ Md Shoaib Khan, **Amit Kumar Gupta** and Manoj Kumar*. BcellEpi: Prediction of linear B-cell epitopes from viruses, bacteria and altered-self (Under communication)

Posters:

- ✓ Md Shoaib Khan**, Amit Kumar Gupta** and Manoj Kumar*. Comparative analysis of immunogenic viral epitopes with altered-self and bacterial epitopes restricted by HLA-A26/B8: Therapeutic implication for viruses. Proceedings of VIROCON 2015, XXIV National Conference of Indian Virological Society (IVS) at NEIGRIHMS, Shillong, India
- ✓ Md Shoaib Khan**, Amit Kumar Gupta** and Manoj Kumar*. Prediction and analysis of viral and bacterial MHC class II immunogenic and non-immunogenic epitopes restricted to human host. INTERVIROCON 2018, International Conference of Virology, Global Viral Epidemics: A Challenging Threat, PGIMER, Chandigarh, India

SCIENTIFIC REPORTS

OPEN

Received: 04 November 2014
Accepted: 26 June 2015
Published: 24 July 2015

HPVbase – a knowledgebase of viral integrations, methylation patterns and microRNAs aberrant expression: As potential biomarkers for Human papillomaviruses mediated carcinomas

Amit Kumar Gupta & Manoj Kumar

Human papillomaviruses (HPVs) are extremely associated with different carcinomas. Despite consequential accomplishments, there is still need to establish more promising biomarkers to discriminate cancerous progressions. Therefore, we have developed HPVbase (http://crdd.osdd.net/servers/hpvbase/), a comprehensive resource for three major efficacious cancer biomarkers i.e. integration and breakpoint events, HPVs methylation patterns and HPV mediated aberrant expression of distinct host microRNAs (miRNAs). It includes clinically important 1257 integrants and integration sites from different HPV types i.e. 16, 18, 31, 33 and 45 associated with distinct histological conditions. An inclusive HPV integrant and breakpoints browser was designed to provide easy browsing and straightforward analysis. Our study also provides 719 major quantitative HPV DNA methylation observations distributed in 5 distinct HPV genotypes from higher to lower in numbers namely HPV 16 (495), HPV 18 (113), HPV45 (66), HPV 31 (34) and HPV 33 (11). Additionally, we have curated and compiled clinically significant aberrant expression profile of 341 miRNAs including their target genes in distinct carcinomas, which can be utilized for miRNA therapeutics. A user-friendly web interface has been developed for easy data retrieval and analysis. We foresee that HPVbase an integrated and multi-comparative platform would facilitate reliable cancer diagnostics and prognosis.

The human papillomaviruses (HPVs) are circular, double-stranded DNA genome of approximately 8.0 kb in length. It belongs to the papillomaviridae family, which is further taxonomically classified into distinct genera namely alpha, beta, gamma, mupa and nupa[1]. HPVs encode eight well-defined open reading frames (ORFs) along with one non-coding long control region (LCR) or regulatory region. HPV proteins are mainly divided into two coding regions classified as early (E) and late (L) region. E region includes six ORFs encoding 3 functional regulatory genes (E1, E2, E4), 3 oncogenes (E5, E6, E7) and the L region encodes the two viral capsid genes (L1 and L2)[1–3].

Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research (CSIR), Sector 39-A, Chandigarh; 160036, India. Correspondence and requests for materials should be addressed to M.K. (email: manojk@imtech.res.in)

Genomics 112 (2020) 4853–4862

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

HPVomics: An integrated resource for the human papillomavirus epitome and therapeutics

Amit Kumar Gupta[a], Manoj Kumar[a,b,*]

[a] Virology Unit and Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research (CSIR), Sector 39-A, Chandigarh 160036, India
[b] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

ARTICLE INFO

Keywords:
HPVs
Oncogenes
Therapeutics
Vaccine epitopes
Algorithm, database

ABSTRACT

Human papillomaviruses (HPVs) belongs to the *Papillomaviridae* family, which is divided into high-risk (HR), and low-risk (LR) HPVs based on their disease-causing competence. HR-HPVs 16 and 18 are known to cause distinct carcinomas like cervical and head and neck, whereas LR-HPVs are commonly associated with the genital warts. We have developed an integrative platform; HPVomics dedicated to the potential therapeutic regimens targeting all HPV genes including oncoproteins E6, E7 and E5. We primarily focused on eighteen HR-HPVs and eleven LR-HPVs. It mainly deals with therapeutically imperative elements, i.e., vaccine epitopes, siRNAs, sgRNAs, and antiviral peptides. Simultaneously, it also comprises of genome browser, whole-genome sequences and annotation of HPVs with searching and filtering capabilities. Moreover, we have also developed an integrated support vector machine (SVM) based computational algorithm "*HPVepi*" for the prediction of HPV epitome. We hope that HPVomics (http://bioinfo.imtech.res.in/manojk/hpvomics/) will assist the scientific community engaged in HPV research.

1. Introduction

Human papillomaviruses (HPVs) are circular, double-stranded DNA viruses belong to the *Papillomaviridae* family, which are known to cause different carcinomas and genital warts [1,2]. These are approximately 8.0 kb in length that encodes eight well-defined open reading frames (ORFs) along with one long control region (LCR) or regulatory region. HPVs genome is broadly divided into two coding sections (1) early (E) gene region that includes six ORFs encoding 3 functional regulatory genes (E1, E2, E4), and 3 oncogenes (E5, E6, E7); (2) late (L) gene region comprise of two ORFs, i.e., L1 and L2, which encode two viral structural proteins viz. major and minor capsid proteins, respectively. HPVs are classified into distinct genera namely *Alpha*, *Beta*, *Gamma*, *Mu*, and *Nu* [3,4].

HPVs are known to infect mucosal and cutaneous epithelial tissues. According to the malignant transformation competence, these are classified into distinct subgroups: high-risk HPVs (HR-HPVs highly carcinogenic) associated to diverse cancers and low-risk HPVs (LR-HPVs) which are mainly linked with genital warts. Persistent HR-HPVs infection is associated with the cancer progression and can cause diverse array of malignancies, i.e., cervical, oropharyngeal, penile, vulvar and anal carcinomas [1,5–7]. HR-HPV types usually 16 and 18 are prevalent in the etiology of human carcinomas and play a cardinal role in the cervical cancer, which is the fourth most common cancer in women [8–11].

In the HPV carcinogenesis, E6 and E7 oncoproteins are considered as the most preferred and ideal target for the therapeutic vaccines as they play a crucial part in the HPV mediated malignant transformations, i.e., from low-grade cervical intraepithelial neoplasia (CIN 1) to high-grade CIN 2/3 and finally into invasive cervical cancer (ICC) [12–14]. E6 and E7 protein degrade the p53 (apoptosis regulator) and the tumor suppressor retinoblastoma protein (pRb), respectively, which further disrupt the cell apoptosis, and cell cycle regulation. This leads to the abnormal cell growth, host genomic instability and eventually cancer progression [1,2,15–17]. Along with this, E5 is also considered as an oncogene and several researches also suggest their role in HPV carcinogenesis [18,19].

HR- and LR-HPV types are most critical and bear a priority in terms of vaccine development against them. Several efforts are made to prevent HPV induced diseases by employing prophylactic and immunotherapeutic vaccine approaches [15,20]. Three prophylactic vaccines based on the HPV L1-virus like particles (VLPs) have been developed and approved to resist HPV infections [21,22]. Earlier, two HPV vaccines, a quadrivalent HPV-6/11/16/18 vaccine named as

* Corresponding author at: Virology Unit and Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research (CSIR), Sector 39-A, Chandigarh 160036, India.
E-mail addresses: amitg@imtech.res.in (A.K. Gupta), manojk@imtech.res.in (M. Kumar).

https://doi.org/10.1016/j.ygeno.2020.08.025
Received 6 December 2019; Received in revised form 7 August 2020; Accepted 19 August 2020
Available online 29 August 2020
0888-7543/

Table of Contents

Introduction

Chapter 1. Introduction

PART-I

Viruses and global disease burden

Viruses are the known most abundant bodies on the earth. They are the pathogenic agents of various diseases and affect the livelihood of millions of people worldwide (Virgin, 2014). Viruses such as Influenza, Hepatitis, Nipah, Zika, Ebola, Corona, etc. are highly infectious and can be a cause of deadly outcomes (Virgin, 2014). Further, the outbreak and epidemic of these emerging and remerging viruses make the situation even worse. This imparts a great socio-economical burden and hinders growth and development. Further, Bacteriophages (bacteria-killing viruses) play a critical role and influence natural ecosystems (environments) and human microbiota (Abeles and Pride, 2014; Minot et al., 2013; Paez-Espino et al., 2016). They are also known vehicles for the transmission of antibiotic resistance genes (ARGs), transposable elements, genetic materials (transduction) (Abeles and Pride, 2014; Balcazar, 2014). Moreover, viruses also play a cardinal role in the establishment and progression of different carcinomas, which leads to the high mortality rate (Javier and Butel, 2008; Moore and Chang, 2010; Morales-Sánchez and Fuentes-Pananá, 2014).

Oncoviruses and cancers

Viruses that are involved and able to drive cancer are known as oncoviruses (Krump and You, 2018; zur Hausen, 1991). The virus genes which can transform a normal cell into the cancerous and mainly account for oncogenicity are termed as (viral) oncogenes. Further, altered host genes can also promote tumor. Around seven viruses are primarily notorious to cause cancer and significantly account for 10-15 percent of the global cancer burden (Javier and Butel, 2008; Martin and Gutkind, 2008; Moore and Chang, 2010). The event of transmission of tumors from one dog to another is first reported in 1876 from Russia. After three decades, in 1908, Ellerman and Bang have demonstrated that leukemia cell extract from chicken can induce cancer probably due to the transmission of sarcoma leukosis virus. Simultaneously, in 1909, Rous has shown that sarcoma extract from chicken can cause a tumor. Further, Shope identified and reported the first mammalian oncovirus, i.e., cottontail rabbit papillomavirus (also known as Shope papillomavirus) in 1933. Later in 1964, Epstein,

Achong, and Barr discovered the first human tumor virus, a herpesvirus also known as human herpesvirus 4 (HHV4) or Epstein-Barr virus (EBV) from Burkitt lymphoma cells (Moore and Chang, 2010). Further in the 1980s, the hepatitis B virus (HBV) and human papillomavirus (HPV) were identified to be linked with hepatocellular carcinoma (HCC) and cervical cancer, respectively. Both DNA and RNA viruses are capable of inducing and contribute towards the advancing of distinct carcinomas (Mesri et al., 2014; Mui et al., 2017) (**Table 1**).

Table 1. List of different oncoviruses and associated cancer/diseases

Viruses	**Type**	**Cancer or disease**
Human herpesvirus 4 (HHV4) (also known as Epstein-Barr virus (EBV))	dsDNA	Burkitt`s lymphoma Hodgkin`s lymphoma Nasopharyngeal carcinoma
Hepatitis B virus (HBV) Hepatitis C virus (HCV)	DNA RNA	Hepatocellular carcinoma (Liver cancer)
Human T-cell lymphotropic virus type 1 (HTLV-1)	RNA	Adult T-cell leukemia
Kaposi`s sarcoma-associated herpesvirus (KSHV) (Formally, human herpesvirus 8 (HHV8))	dsDNA	Kaposi`s sarcoma
Merkel cell polyomavirus (MCPyV)	dsDNA	Merkel cell carcinoma
Human papillomaviruses (HPVs)	dsDNA	Cervical, HNSCC, Penile, Vulvar, Anal

In viral carcinogenesis, the sequence of complex molecular mechanisms is usually allied. This mainly includes disruption of cellular DNA damage repair system and cell cycle, genetic and epigenetic abnormalities, viral DNA integration in host genome, inflammation, abrupt dysregulation of genes and microRNAs (miRNAs), inhibition of tumor suppressor proteins. Moreover, oncogenes generally alter and disturb genomic stability, homeostasis, cellular signaling, apoptosis, and immune responses. Eventually, an unavoidable, excessive proliferation of cells occurs that leads to cancer progression and metastasis (Krump and You, 2018; Morales-Sánchez and Fuentes-Pananá, 2014; zur Hausen, 1991).

Human Papillomaviruses (HPVs) and genome organization

Human papillomaviruses (HPVs) are the double-stranded (dsDNA) circular oncovirus from the family *Papillomaviridae*. HPVs genomes are ~8 kb in length and encode 8 open reading frames (ORFs), principally divided into two coding regions, i.e., early (E) and late (L). E region comprises six open reading frames (ORFs) namely E1, E2, E4 (functional regulatory genes), and three oncogenes viz. E5, E6, and E7. Further, the L region forms two capsid genes L1 and L2. Correspondingly, it also encompasses one non-coding regulatory region also recognized as long control region (LCR) (**Figure 1**). HPVs are broadly categorized into the five distinct genera, i.e., *Alpha*, *Beta*, *Gamma*, *Mu*, and *Nu* (de Villiers et al., 2004; Doorbar, 2006; Doorbar et al., 2012). Genome organization is shown in **Figure 1** employing the CGView server (Grant and Stothard, 2008).

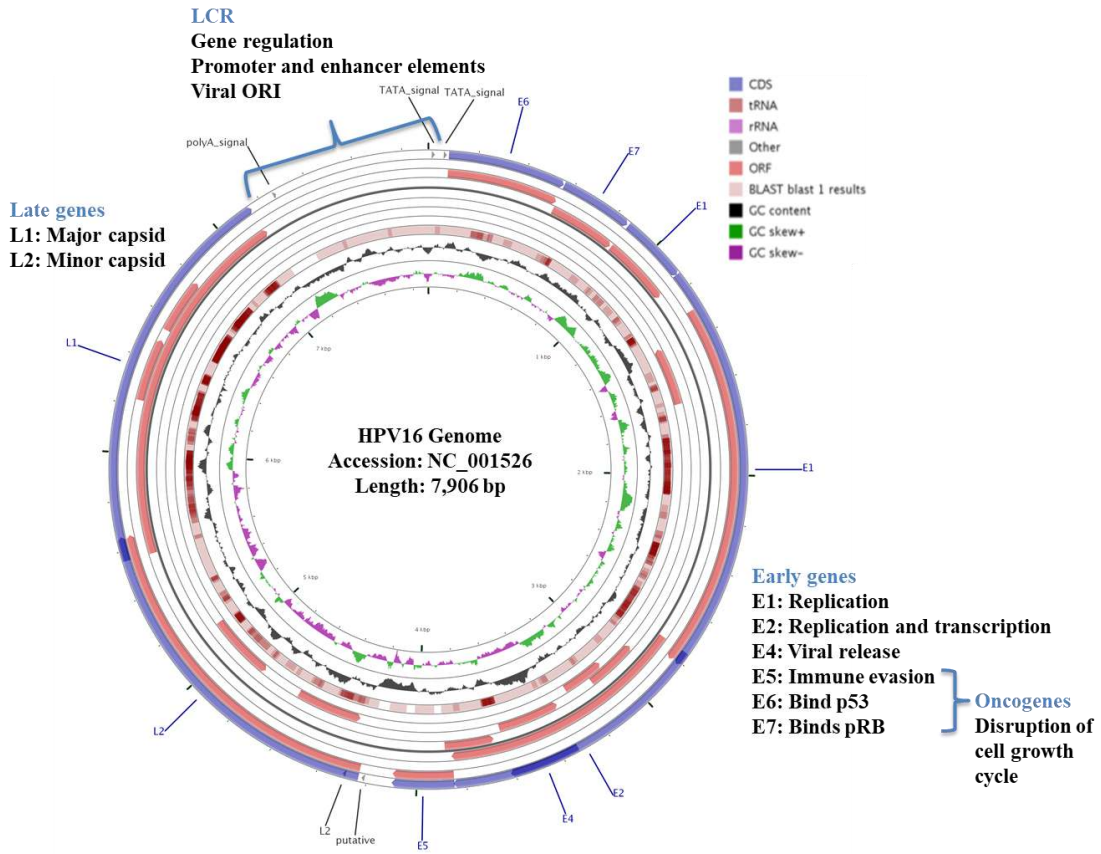


Figure 1. HPV16 genome organization along with gene functions, ORFs and GC content

Biology of HPVs and pathogenesis

The complete life cycle of HPVs contingent on the differentiation of cellular epithelial cells. During HPVs life span, eight HPV genes (E1, E2, E4, E5, E6, L1, L2) use to be expressed at a distinct period thus named as early and late. HPV virions infect mucosal or cutaneous epithelial cells. They enter through micro-abrasions and infect keratinocytes in the basal layer of stratified squamous epithelia (**Figure 2**) (Doorbar, 2006). HPV genome exists and is maintained as episome with 20-100 copies. After entry into the nucleus of the host cell, two early genes, E1 and E2 mediate the virus genome replication and then form the messenger RNAs (mRNAs). Correspondingly, E4 and E5 genes provide support to the replication process. Also, transcription cascading occurs as the cellular cell start differentiation. Most importantly, two oncogenes E6 and E7 play a cardinal role in viral pathogenesis and oncogenicity. E7 binds and inhibits the human retinoblastoma (Rb) gene which is responsible for regular cell division and cell cycle arrest through employing essential proteins like E2F. Likewise, E6 interferes with the functioning of p53 (tumor suppressor gene) and degrade it via the E6AP ubiquitin proteolytic pathway that halts cell damage repairing, abrogates apoptosis, and then finally leads to cancer. Further, L1 and L2 genes form viral capsid proteins for the packaging and generation of new viruses. HPV utilize cellular machinery for survival (Mantovani and Banks, 2001; Moody and Laimins, 2010; Munger et al., 2001; zur Hausen, 2002).

Role of HPVs in cancers-High risk and low-risk HPVs

Based on the tissue tropism and malignancy competence, HPVs are grouped into two categories viz. mucosal and cutaneous types. Mucosal HPVs are further divided into two subgroups i.e. high-risk HPVs (HR-HPVs highly carcinogenic) and low-risk HPVs (LR-HPVs). Persistence infection of these viruses is critical and decisive in the advancement of cancer. Two HR-HPVs, HPV 16 and 18 are predominant in the etiology of human carcinomas and involve in a diverse set of malignancies such as cervical, head and neck squamous cell carcinoma (HNSCC), penile, anal, vulvar, etc. HPV is one of the most common sexually transmitted infection (STI). These are extremely associated with cervical carcinogenesis (Crosbie et al., 2013; zur Hausen, 2002, 2009).

Cervical carcinoma (CaCx) is the fourth utmost cancer and prevailing behind the deaths among women worldwide (GLOBOCAN 2018) (Arbyn et al., 2020; Bray et al., 2018). HPVs are a critical factor in cervical precursor lesions and cancer (Woodman et al., 2007). There are various steps generally involved in cervical oncogenesis, this mainly starts with the HPV infection and transmission, followed by viral persistence, that allows further progression towards precancer or high-grade precursor lesions and invasive carcinoma (**Figure 2**). Histologically and based on severity, cervical cancer is categorized into cervical intraepithelial neoplasia (CIN) I, II, and III (de Villiers et al., 2004; Doorbar, 2006; Doorbar et al., 2012; Munoz et al., 2003; Schiffman and Wentzensen, 2013; Smith et al., 2007). In 2008, Dr. Harald Zur Hausen received the Nobel prize in physiology and medicine for discovering the role of HPVs in cervical cancer (zur Hausen, 2009).

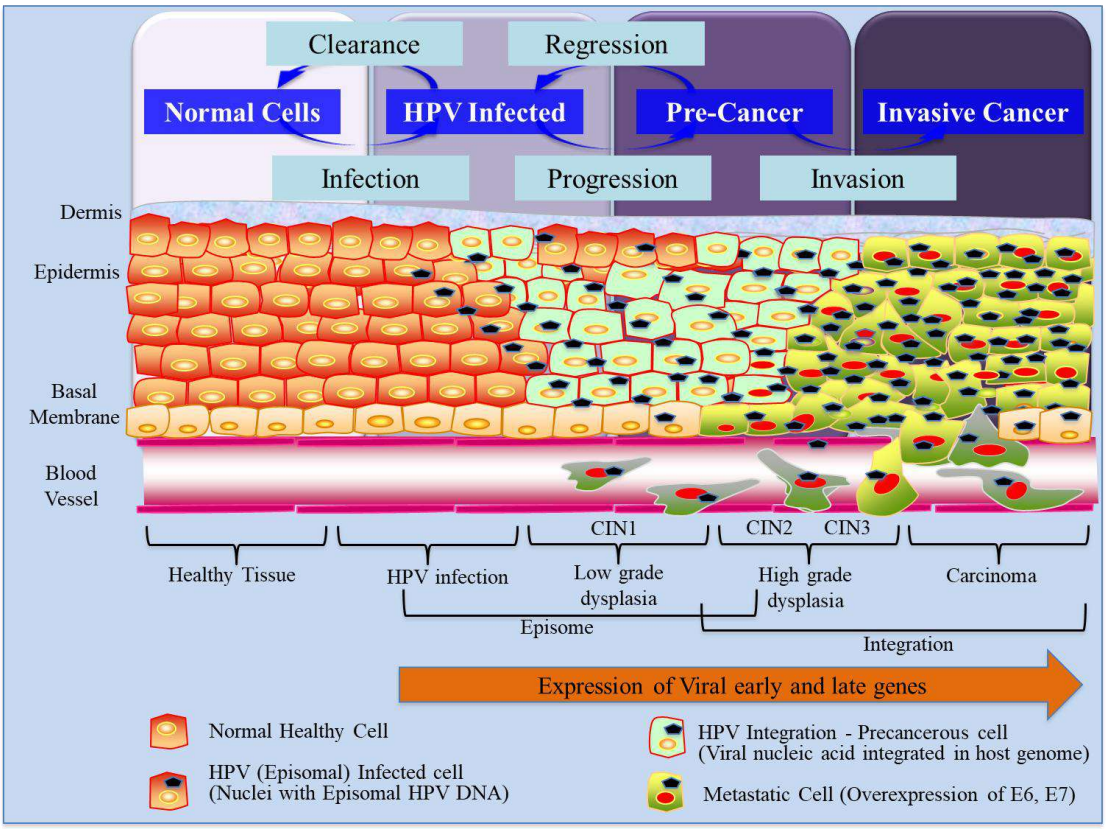


Figure 2. HPV pathogenesis and cancer progression

Vaccine and therapeutic strategies against HPVs

HR-HPV types are most critical and bear a priority in terms of vaccine development against them. As per the report, certain HPV types 16, 18, 31, 33, and 45 are considered as most carcinogenic and account for ~90 % cases of cervical tumors (de Sanjose et al., 2010). Several efforts are made to prevent cervical cancer by employing prophylactic and immunotherapeutic vaccine development approaches. Furthermore, immunotherapy with the peptide-based vaccine has generated new possible ways for the treatment of HPV directed carcinogenesis. It majorly applied and focused on the prevention of cervical cancer (Chabeda et al., 2018; Cheng et al., 2018; Dadar et al., 2018; Smith et al., 2005).

To date, distinct epitopes of HPVs (primarily for HPV16, 18) especially for the E6 and E7 oncogenic proteins are identified and reported. The two non-structural oncoproteins E6 and E7 are considered as the most preferred and ideal target for the therapeutic vaccines as they consistently expressed, play a crucial part in the HPV mediated malignant transformations (Devaraj et al., 2003; Jansen and Shaw, 2004; Morrow et al., 2013; Roden et al., 2004). These are largely discussed and diagnosed concerning cervical cancers (Bourgault Villada et al., 2000; Facciuto et al., 2014; Kather et al., 2003; Liu et al., 2007; Morishima et al., 2007; Nakagawa et al., 2004; Riemer et al., 2010; Rudolf et al., 2001; Youde et al., 2000). Along with this, HPV E5 is also considered as an oncogene and several researchers also suggest their role in carcinogenesis as it can alter distinct cellular pathways (Chang et al., 2001; Kim et al., 2010; Liao et al., 2013a; Liao et al., 2013b; Maufort et al., 2010). Additionally, some studies also identified HPV-16 E5 protein as a tumor rejection antigen and demonstrate the E5 as target antigens (DiMaio and Petti, 2013; Liu et al., 2000).

The majority of HPV infections are known to be short-term or may be cleared by the host immune system (Frazer, 2004) and mostly depend on strong cell-mediated immune responses. The significance of cell-mediated immune response in eradicating HPV infections is well known and reported (Einstein et al., 2009; Grabowska and Riemer, 2012). Specifically, for the clearance of persistent HPV infections antigen-specific T cell-mediated immunity is substantial (Stanley, 2006; Testa and Philip, 2012). Various studies have clinically tested DNA vaccines, protein vaccines, and CTL epitopes from E6 and E7 of HPV16-18, nevertheless are unable to exhibit

promising clinical efficacy (Einstein et al., 2007; Galloway, 2003; Garcia-Hernandez et al., 2006; Hallez et al., 2004; Kaufmann et al., 2002; Peng et al., 2007; Ressing et al., 1995; Steller et al., 1998). Additionally, it is also determined that HPV therapeutic vaccines are not able to completely eradicate the lesions (Peng et al., 2006; Sarkar et al., 2005). However, for the initiation of HPV-specific T lymphocytes, vaccination is a promising approach. In past, a study has shown efficacious clinical outcome describing a polyepitope vaccine created using multiple long synthetic peptide fragments of E6 and E7, which was tested on HPV-16+ vulvar intraepithelial neoplastic patient (Kenter et al., 2009).

Furthermore, prophylactic vaccines utilizing virus-like particle (VLP) consist of L1 capsid proteins was also established to resist HPV-induced malignancy. Till date, three prophylactic vaccines were developed. Previously, a quadrivalent recombinant vaccine named Merck`s Gardasil for HPV 6; 11; 16 and 18, GlaxoSmithKline`s Cervarix, a bivalent vaccine against HPV 16 and 18 was developed (Berzofsky et al., 2004; Descamps et al., 2009; Harper, 2009; Keam and Harper, 2008; Lowy and Schiller, 2006; Paavonen et al., 2009; Siddiqui and Perry, 2006; Tjalma et al., 2004). These are mostly known to prevent HPV 16 and 18 infections only (Govan, 2008). Recently Gardasil 9, a novel recombinant nonavalent vaccine targeting 9 HPV types i.e. 6, 11, 16, 18, 31, 33, 45, 52, and 58 is developed to prevent HPV infection (Huh et al., 2017; Joura et al., 2015). However, due to the late expression of capsid genes during replication, these vaccines are not capable of effectively abolishing established viral infections (Hildesheim et al., 2007; Hu and Ma, 2018). Further, these remain ineffective because of non-productive infection of HPVs in which viral capsid proteins remains unexpressed and ultimately generate poor clinical retort (Chabeda et al., 2018; Dadar et al., 2018; Frazer, 2004; Longworth and Laimins, 2004; Munger et al., 2004). However, for the development of effective therapeutic vaccines, numerous approaches utilizing nucleic acid, peptide or protein, live-vectors, etc. individually or in combination are under investigation and in clinical trials (Chabeda et al., 2018; Cheng et al., 2018; Dadar et al., 2018).

Key events during HPV infection to carcinoma

HPV infection is very common in women. However, only a few uses to progress into invasive high-grade carcinoma, i.e., CIN2 and CIN3 (Cogliano et al., 2005; Schiffman et al., 2007; Schiffman et al., 2011). With the active participation and influence of

HPVs, research on screening and prevention strategies for cancer is accelerated. Correspondingly, there is a worldwide reduction in the cancer morbidity and mortality rate with the advent of HPV-based screening and cytology approaches. (Boulet et al., 2008). Though, deciding to refer for a colposcopy test is still challenging due to the irregular sensitivity and specificity of these methods. Further, there is an inadequacy in identifying and differentiate transient infections that have a higher tendency to stride towards persistence infection or precancer and high-grade cervical carcinomas (Clarke et al., 2012; Crosbie et al., 2013; Woodman et al., 2007).

However, consequences or key events allied with the HPV infection are a valuable aspects of multistage cancer development (Schiffman and Wentzensen, 2013). These can be looked at as an alternative biomarker to differentiate latent infection with high-grade precursor lesions and cancer. These HPV related factors mainly include viral DNA integration, epigenetic modifications (methylation), aberrant expression of microRNAs (miRNAs) and HPV variants and mutations (heterogeneity) along with other cellular and environmental risk factors (**Figure 3**) (Sahasrabuddhe et al., 2011; Schiffman and Wentzensen, 2013).

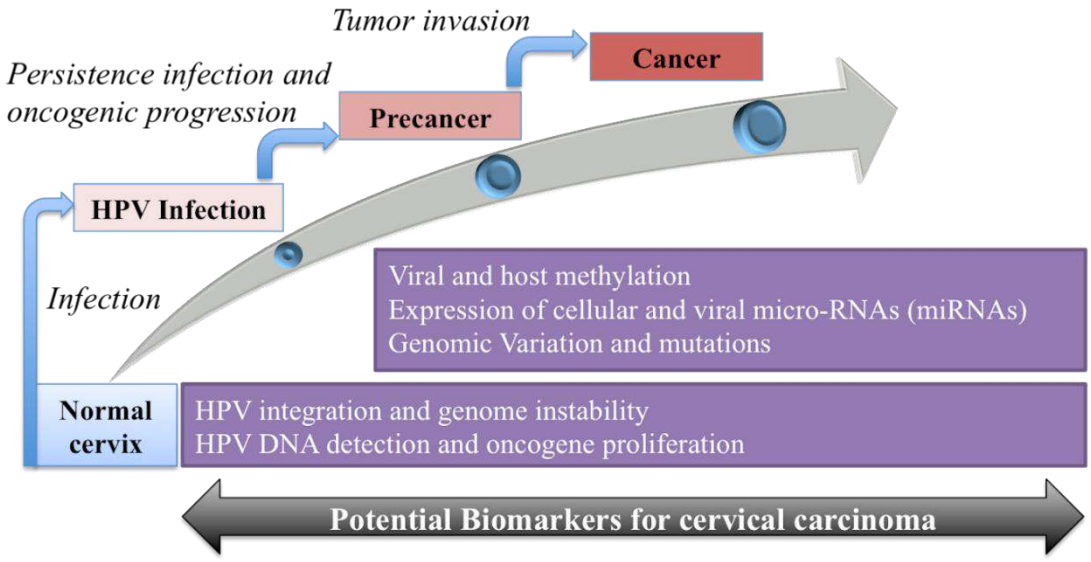


Figure 3. Multistage neoplastic progression and associated events

Viral Genome Integration

After HPV infection and persistency, the integration of viral DNA into the cellular genome is a well-known and crucial phenomenon in the etiology and progression of different carcinomas. This is highly associated with Cervical and head and neck

carcinoma (Bosch et al., 2002; Gillison et al., 2000; Gillison and Shah, 2003; McBride and Warburton, 2017; Oyervides-Muñoz et al., 2018; Schwarz et al., 1985; Walboomers et al., 1999; Zandberg et al., 2013; zur Hausen, 2002). Further, the incidence and frequency of integration increase with cancer severity in high-grade lesions and carcinomas with irregular occurrence in different cases (Wentzensen, Vinokurova et al. 2004)(Matovina et al., 2009; McBride and Warburton, 2017; Schiffman and Wentzensen, 2013). However, HPVs usually remain in the episomal form in benign and low-grade lesions (Klaes et al., 1999; McBride and Warburton, 2017; Schiffman and Wentzensen, 2013; Williams et al., 2011).

Moreover, Integration events are known to provide stability and enhancement to E6 and E7 (HPV oncogenes) expression during the progression of cancer (Jeon and Lambert, 1995; McBride and Warburton, 2017; Oyervides-Muñoz et al., 2018; Schiffman and Wentzensen, 2013; Wentzensen et al., 2004). Although these integrations happen randomly and scattered all over the host genome, evidentiary reports suggesting that there are an abundance and preference for the integration, in or near to oncogenes, translocation breakpoints and fragile sites may be due to the higher genomic instability or larger chromosomal deletions and rearrangements nature (McBride and Warburton, 2017; Schiffman and Wentzensen, 2013; Thorland et al., 2003). It is also proposed that certain specific integrations lead or contribute to malignant transformation and tumor development (Parfenov et al., 2014; Schiffman and Wentzensen, 2013).

Fragile sites (FS) are specific regions in the genome that show higher instability and are especially prone to gaps or breaks on metaphase chromosomes in response to stress (McBride and Warburton, 2017; Yunis and Soreng, 1984). It also plays an influential role in cancer-specific translocations, gene rearrangements. Many tumor suppressors and oncogenes tend to locate and identified within or near fragile sites. Breakages at these sites often lead to the deletion of tumor suppressors and enhanced oncogenic amplifications (Durkin and Glover, 2007; McBride and Warburton, 2017; Yunis and Soreng, 1984). These regions are the common targets for the different oncogenic viruses to integrate. Integrations within or near to common fragile sites are usually associated with frequent deletions and instability that are frequently observed in several tumor types (Cannizzaro et al., 1988; Matovina et al., 2009; McBride and Warburton, 2017; Thorland et al., 2003).

Moreover, HPV DNA integration into the host genome leads to the disruption of the episomal state and results in the deletion of HPV E1/E2 repressor regions that further enhance transcription of the oncogenic E6/E7 proteins (Choo et al., 1988; McBride and Warburton, 2017; Oyervides-Muñoz et al., 2018; Xue et al., 2010). Higher and continuous the expression of these oncoproteins govern malignancy by binding and inactivating the p53/Rb tumor suppressor proteins of host normal cell cycle with other alterations (Dyson et al., 1989; McBride and Warburton, 2017; Oyervides-Muñoz et al., 2018; Scheffner et al., 1990; Werness et al., 1990). In another study, Akagi and Li et al. used whole-genome sequencing and high throughput molecular techniques to shed new light on the association between HPV integration and genome instability regions such as sites of amplification and rearrangements. They utilize 10 HPV positive and negative human cancer cell lines (5-5 respectively), which were derived from cervical and head and neck cancer. Authors proposed a model of "looping" based on the finding that host genomic instability sites are flanked and bridged by HPV integrant, which results in viral-host DNA concatemers, and leads to the oncogene disruption (Akagi et al., 2014). However, complete knowledge and understanding of the integration mechanism are still unclear.

Dysregulation of miRNAs

The small endogenous non-coding regulatory RNA species, microRNAs (miRNAs) are significant post-transcriptional regulators of gene expression. MiRNAs attaches to the 3` UTRs of target messenger RNAs (mRNAs) and often negatively regulate their mRNA targets (Kim, 2005a). Mature miRNAs are roughly 18-25 nucleotides (nts) in length, non-immunogenic, and present in a variety of organisms including viruses (Kim, 2005c). They can catalyze mRNA cleavage or suppress translation to alter gene expression (Engels and Hutvagner, 2006; Hutvagner and Zamore, 2002; Lee et al., 2008a; Valencia-Sanchez et al., 2006). miRNAs are known to influence numerous molecular mechanisms mainly cell proliferation, gene expression, morphogenesis, apoptosis, chromatin modification, and tumorigenesis (Arlotta and Macklis, 2005; Iorio and Croce, 2012; Kim, 2005a, c; Winter and Diederichs, 2011; Winter et al., 2009). However, most functions of miRNAs are yet undetermined.

Due to high implications, the biogenesis of miRNAs is exceedingly important. There are two ways for the formation of precursor-miRNA hairpins (pre-miRNAs) usually 60-110 nucleotides in length. Either mediated by the Drosha, an RNase-III enzyme

responsible for the production of pre-miRNA from independent primary miRNA transcripts (pri-miRNAs) in complex with DGCR8; a double-stranded RNA-binding protein also known as DiGeorge syndrome critical region gene 8 (Han et al., 2004; Han et al., 2006; Wang et al., 2007) or by reprogramming of mRNA introns through cis/trans-splicing events (Ruby et al., 2007). Further, exportin 5 intercede transfer of pre-miRNAs from nucleus to the cytoplasm, where cleavage activity of another RNase III enzyme known as Dicer will produce miRNA duplex (Gregory et al., 2005; Yi et al., 2003) and one strand (known as a guide strand) of duplex form mature miRNAs by the action of the RNA-induced silencing complex (RISC). The remaining strand (referred to as miRNA-star) is often considered to be degraded or non-functional (Schwarz et al., 2003). However, an alternative pathway for the human miR-451 was also stated, in which Argonaut 2 (AGO2) cleaves pre-miRNAs to form AGO2-cleaved pre-miRNAs (ac-pre-miRNAs) and further processed to mature miRNAs by exonucleolytic digestion (Bartel, 2004; Cheloufi et al., 2010; Cifuentes et al., 2010; Diederichs and Haber, 2007).

The aberrant (increased or decreased) expression of miRNAs is evidentiary in diverse cancers and viral diseases (Farazi et al., 2011; Gocze et al., 2013; Lui et al., 2007; Rao et al., 2012; Wang et al., 2008; Wang et al., 2014; Wilting et al., 2013). They are located near or at genomic unstable regions such as chromosomal fragile sites, at genomic rearrangements and amplification regions (Calin et al., 2004; Gomez-Gomez et al., 2013; Thorland et al., 2003). Moreover, also observed at HPV integration sites and found associated with clinical outcomes of cancers (Calin and Croce, 2006; Calin et al., 2004; Zhang et al., 2006). In addition to the HPV mediated genomic instability and epigenetic regulations, the mechanism of cancer transformation is even more complicated. The two HPV cancer-causing genes E6, and E7 interact with various cellular genes, modulate transcriptional regulation and may alter distinct cellular pathways and mechanisms (Moody and Laimins, 2010; Yim and Park, 2005). These oncogenes tend to deregulate the oncogenic miRNAs (oncomirs) as well as tumor-suppressive miRNAs (Esquela-Kerscher and Slack, 2006; Gomez-Gomez et al., 2013; Reshmi and Pillai, 2008). Moreover, cellular miRNAs are also known to target viral RNA transcripts hence affect the expression of HPV genes accordingly run back-and-forth mechanism (Zheng and Wang, 2011).

In various studies, miRNAs are suggested being involved in the pathogenesis of cancers usually in cervical and head and neck carcinomas (Hu et al., 2010; Martinez et al., 2008; Wald et al., 2011; Wang et al., 2008) including other cancer types, i.e., breast cancer (Iorio et al., 2005), thyroid papillary malignancy (Pallante et al., 2006; Visone et al., 2007), ovarian cancer (Iorio et al., 2007), prostate cancer (Nadiminty et al., 2012; Ozen et al., 2008; Zaman et al., 2010), etc. Numerous literature reported significantly overexpressed miRNAs that may act as oncogenes (Oncomirs) in cervical carcinomas; including hsa-mir-16, 21, 25, 92a, 127, 135b, 146a, 182, 199a, 205, 223, 224, 301b, 378 etc. (Lee et al., 2008a; Shen et al., 2013; Tang et al., 2013; Wang et al., 2008; Wang et al., 2014; Xie et al., 2012). whereas, previous studies also report highly downregulated miRNAs as tumor suppressors such as has-mir- 22, 27a, 29a, 34a, 100, 124, 143, 145, 200a, 214, 218, 433 etc. (Hu et al., 2010; Lajer et al., 2012; Li et al., 2010a; Lui et al., 2007; Martinez et al., 2008; Pang et al., 2010; Wang et al., 2008; Wang et al., 2014; Wang et al., 2009; Wilting et al., 2010; Yamamoto et al., 2013; Yang et al., 2009; Zhou et al., 2010).

Along with cellular miRNAs, several DNA/RNA viruses also encode their miRNAs. Viral miRNAs can regulate both viral as well as cellular mRNAs (Kincaid and Sullivan, 2012; Murphy et al., 2008; Pfeffer et al., 2004). Potentially, viral miRNAs modulate cellular programming associated with host immune responses that could be achieved by targeting the cellular genes associated with cell proliferation, host defense mechanism, and immune recognition (Pfeffer and Voinnet, 2006; Skalsky and Cullen, 2010). These processes are very crucial for persistent viral infection and subsequently lead to the enhanced viral expression (Bauman et al., 2011; Lee et al., 2011; Seo et al., 2008). Most likely, viral miRNAs mediate viral survival and provide support for consistent infection by promoting immune evasion thus contributes to cancer development (Pfeffer and Voinnet, 2006).

The commonly known viral miRNAs are found in polyomaviruses (Chen et al., 2011; Lee et al., 2011; Seo et al., 2009; Sullivan et al., 2005), adenoviruses (Xu et al., 2007)), herpesviruses (Amoroso et al., 2011; Besecker et al., 2009; Grundhoff and Sullivan, 2011)), and in ascoviruses; the family of double-stranded DNA viruses (Hussain et al., 2008). A study also identified and reported a few HPV encoded miRNAs and their probable targets with a suggestive role in host cell interactions,

immune regulation, cellular morphology, and oncogenesis (Qian et al., 2013). Another work also provides strong evidence for the presence of viral miRNAs by predicting different HPV types (Gu et al., 2011). These findings guide towards the remarkable possibility for the development of antiviral drugs targeting viral and cellular miRNAs, as they show evocative functions in viral infection and carcinogenesis.

Epigenetic modifications (Viral and host methylation)

Epigenetic modifications play a critical role and alter the conformation of chromatin. These are also known to regulate the expression of genes. It suppresses gene activity by interrupting transcription factor binding sites or by recruiting histone deacetylases through methyl-CpG-specific repressor proteins (Nan et al., 1998; Rountree et al., 2001). DNA methylation is one of the molecular regulatory processes in epigenetic that refers to the covalent addition of a methyl group to cytosine residues intrinsically occurs at CpG dinucleotides (MeCpGs), which mediates binding of MeCpG-specific transcriptional repressors like MeCP2 (Doerfler, 2005; Fuks, 2005; Klose et al., 2005). Molecular methods pertinent to divulge CpG methylation include Southern blotting, PCR, quantitative methylation-specific PCR (Q-MSP), cloning, pyrosequencing, and bisulphite sequencing (Brandsma et al., 2009; Crosbie et al., 2013; Mirabello et al., 2012; Turan et al., 2007). It plays a noteworthy role in carcinogenesis and may expedite interaction between genotype and environment (Lorincz, 2011; Robertson, 2005).

Methylation event also mediates HPVs transcriptional modulation (Clarke et al., 2012). However, the associated molecular elements underlying methylation of specific CpG sites remains elusive (Clarke et al., 2012). There are primarily two fundamental processes that were recognized in the previous studies. First, methylation may block the E2BSs binding sites of HPV E2 repressor, which enhance E6 and E7 expression, which then contributes towards the carcinogenic progression. Second, it can be due to the de novo methylation as a cellular defense process to inhibit the replication and transcription of the integrated viral genome (Brandsma et al., 2009; Crosbie et al., 2013; Mirabello et al., 2012; Stunkel and Bernard, 1999; Turan et al., 2007).

The profound research in many cancers demonstrates prodigious assurance in the quantification of HPV DNA methylation as a prominent diagnostic and prognostic novel biomarker. In several studies, it was shown that HPV (mainly HPV16, 18, 31, 33, 45) methylation occur regularly in vivo in cervical cells, in clinical samples and cell cultures (Brandsma et al., 2009; Mirabello et al., 2013; Vasiljevic et al., 2014; Wentzensen et al., 2012). The quantitation of methylation also demonstrates great promise as a simple test for triage of HPV infected women to colposcopy (Bryant et al., 2014). It may accelerate the diagnosis and prognosis of cancer progression (Tornesello et al., 2013). The integration of HPV genomes in carcinoma usually correlates with elevated DNA methylation. In HPV positive women, the methylation level at specific CpGs increase with the consistent viral infection and even enhance more remarkably in high-grade lesions. In addition, cross-sectional studies also depict similar observations with disease progression but show diverges outcome based on the sample type, detection method/assay used (Badal et al., 2003; Brandsma et al., 2009) (Piyathilake et al., 2011).

Assorted cancer studies have advocated an association between CpG methylation patterns and carcinogenic development. Among all, over methylation of late HPV regions L1/L2 in high-grade lesions are regular and conclusive. It has emerged as a cost-effective molecular tool for the triage of HPV +ve women. Hyper methylation of viral L1/L2 gene is most frequent in carcinomas and increase with the severity but absent or rare in low-grade precancerous lesions or asymptomatic infections (Fernandez et al., 2009; Kalantari et al., 2010; Lorincz et al., 2013; Sun et al., 2011a; Turan et al., 2006). Besides, the methylation pattern in the adjacent long control region (LCR) was relatively inconsistent and contradictory. LCR DNA methylation is the most important in terms of viral gene expression as transcription of viral oncogenes E6 and E7, which is crucial for malignant transformation, rely on promoter and enhancer core regions of LCR. Indeed, E2BSs in HPV LCR are the likely targets for methylation (Kim et al., 2003). Some literature have reported significantly increased methylation of CpG sites within the LCR associated with carcinogenesis (Bhattacharjee and Sengupta, 2006a; Ding et al., 2009) and high-grade lesions though others found hypermethylation in case of asymptomatic and low-grade infections (Badal et al., 2003; Mazumder Indra et al., 2011; Xi et al., 2011). However, careful quest suggests that methylation patterns and rates differ according to pathological

conditions and severities. These changes in methylation profile can be utilized as a predictive biomarker to distinguish HPV infections from those that evolve to the cancerous state (Badal et al., 2004; Brandsma et al., 2014; Ding et al., 2009; Patel et al., 2012; Tornesello et al., 2013).

HPV variants and mutations: role in carcinogenesis

As per Papillomavirus Nomenclature Committee, HPVs can be defined into types, subtypes, and intra-types based on nucleotide sequence variation of more than 10%, between 2-10% and below 2% in coding sequence and 5% in the noncoding region respectively (Burk et al., 2013). Over the long period, they are evolved into multiple ethnic lineages (Bernard et al., 2006; de Villiers, 2013). HPV 16 is the most widespread high-risk HPV type (Clifford et al., 2006; Munoz et al., 2006; Smith et al., 2011; Smith et al., 2007). HPV16 variants are well defined into the distinct evolutionary lineages based on the geographical distribution. These groups are as follows: As (Asian; South-East Asia region), E (European; All regions except Africa), AA (Asian–American; found in Central and South America), Af1 and Af2 (African-1/2; in Africa), NA1 (North American; America) and Java (Javanese; in Indonesia). These are identified and grouped based on E6, L1, L2, and LCR sequence variations (Bernard et al., 2006; Chen et al., 2015; Cornet et al., 2012; de Boer et al., 2004; de Villiers, 2013; Pillai et al., 2009; Xi et al., 1997; Yamada et al., 1997).

HPVs mutate slowly and coexist with human mankind (Ho et al., 1993). Mutations in the HPVs especially within the oncoproteins (E6 and E7), L1, and LCR region are associated with cervical cancer etiology with other environmental factors. These variations are used to provide an understanding of viral oncogenic potential. The progression of HPV infection and variations are associated with each other. Various studies have shown the sequence variations related to cervical cancer in different regions. Moreover, intratype variants are also known to help cancer progression (Lichtig et al., 2006; Londesborough et al., 1996; Pillai et al., 2009; Song et al., 1997; Tornesello et al., 2004; Wu et al., 2006; Yamada et al., 1997; Zuna et al., 2009). The association between HR-HPVs (16/18) variations and oncogenic lesions is reported in various literature (Berumen et al., 2001; Cai et al., 2010; Chan et al., 2002; Chansaenroj et al., 2012; Choo et al., 2000; Ding et al., 2010; Grodzki et al., 2006; Hu et al., 2011; Qiu et al., 2007; Shang et al., 2011; Sichero et al., 2007; Sun et al., 2013; Sun et al., 2011b; Villa et al., 2000; Xi et al., 2007; Xiong et al., 2010).

E6, E7, E2, L1 and LCR regions play a significant role in the cancerous etiology and regulate tumorigenesis (Chakrabarti et al., 2004; Chansaenroj et al., 2012; Eschle et al., 1992; Pande et al., 2008; Stunkel and Bernard, 1999; Sun et al., 2013; Tan et al., 1994; Xi et al., 2017)**.** Mutations in the HPV genomic regions are known to produce alterations in the amino acid sequence of functional domains and eventually modify biological processes. Polymorphism in these regions use to influence gene regulation, host immune responses, pathogenicity, stimulate p53 degradation, enhance promoter activity thus liable to effect carcinogenicity (Bernard et al., 2006; Chansaenroj et al., 2012; Kammer et al., 2002; Mantovani and Banks, 2001; Pientong et al., 2013; Stoppler et al., 1996; Xi et al., 2017).

In HPV E6 oncogenic protein; L83V (T350G) amino acid mutation is highly significant in the cancer progression and found to be associated with the neoplastic transformation (Andersson et al., 2000; Asadurian et al., 2007; Berumen et al., 2001; Chakrabarti et al., 2004; de Araujo Souza et al., 2008; Lee et al., 2008d; Lichtig et al., 2006; Radhakrishna Pillai et al., 2002; van Duin et al., 2000; Zehbe et al., 2001) (Giannoudis and Herrington, 2001; Hu et al., 2011; Kammer et al., 2002; Matsumoto et al., 2000). Additionally, variation at D25E is also described as a most associated site contributing to cervical oncogenicity (Cai et al., 2010; Chan et al., 2002; Kang et al., 2005; Matsumoto et al., 2003; Matsumoto et al., 2000; Nindl et al., 1999; Wu et al., 2006; Yamada et al., 1997; Zehbe et al., 1998). Moreover, E113D mutation was also found to be linked with cervical carcinoma (Picconi et al., 2003; Wu et al., 2006). While the role of a point mutation in oncogene inhibition is also reported in literature such as E6 F47R is known to convert HPV16 oncoprotein into a potential suppressor of cell proliferation (Ristriani et al., 2009).

Another oncogene E7 comparatively shows more conservation than E6 (Chen et al., 2005; Garcia-Vallve et al., 2005; Smith et al., 2011; Yamada et al., 1997). However, very well-known variation (N29S) within the retinoblastoma suppressor protein (pRB) binding domain is liable for the oncogenic effect (de Boer et al., 2004; Duensing and Munger, 2002; Jones et al., 1990; Stephen et al., 2000). Furthermore, another frequently reported E7 variation is S63F (Fujinaga et al., 1994; Wu et al., 2006)(Chan et al., 2002; Eschle et al., 1992; Nindl et al., 1999; Radhakrishna Pillai et al., 2002; Song et al., 1997).

HPV E2 region mainly consists of three domains namely: transactivation domain, hinge domain, and DNA-binding domain. Variations in these domains are known to promote viral persistence and upregulation of oncogenic (E6/E7) proteins (Bhattacharjee and Sengupta, 2006b). T310K mutation in E2 shows a correlation with the high-grade lesions (Cornet et al., 2012; Giannoudis et al., 2001; Soeda et al., 2006). Similarly, mutations in L1 protein usually found near viral immunodominant regions and are prone to effect viral antigenicity and may change the epitope conformation (Chansaenroj et al., 2012; Shang et al., 2011; Villa et al., 2000; Wu et al., 2006).

LCR is one of the most important segments of HPVs. It is responsible for the viral transcriptional regulation. It is found to be the uttermost variable region of the HPV16 genome (Hubert, 2005; Kammer et al., 2002; Kammer et al., 2000; Kurvinen et al., 2000; Pande et al., 2008; Schmidt et al., 2001; Shang et al., 2011; Villa et al., 2000; Yamada et al., 1997). Variations in the LCR, specifically within transcription factor binding sites are highly associated with the cancer transformation (Giannoudis and Herrington, 2001; Hubert, 2005; Pientong et al., 2013). Polymorphism in the HPV16 LCR region is suggested as one of the factors in enhancing the expression of the viral oncogene. These are commonly found in relation with the YY-1, TEF-1, SP-1, OCT-1, GRE-1 binding sites and responsible for higher transcription activity (Chen et al., 1997; Dong et al., 1994; Kammer et al., 2000; Kozuka et al., 2000; Pientong et al., 2013; Veress et al., 2001; Veress et al., 1999). Along having this, other variations like insertions, deletions are also associated with E2BSs within the LCR region (Kammer et al., 2000). Moreover, Mutations of R10G/L83V in E6 and the C7294T co-variation in LCR are highly associated with high-grade carcinomas (Sun et al., 2013).

Viruses specific key resources

Along with various experimental accomplishments, numerous computational resources were also developed worldwide to assist in viral research. Some of the resources are as follows. NCBI viral genomes resource provides virus genome sequences and annotations (Brister et al., 2015). ViralZone, a knowledge resource is established to understand virus diversity, virus replication cycle, host-virus interactions, and virion structures (Hulo et al., 2011; Masson et al., 2013). Another, Virus Pathogen Database and Analysis Resource (ViPR) deliver sequence records, gene and protein annotations, 3D protein structures, and visualization tools (Pickett et

al., 2012). Virus Variation Resource for value-added viral sequence data is also designed (Brister et al., 2014). Likewise, viruSITE, an integrated database provides viral genomes and contains information on virus taxonomy, host range, genome features, and viral genes and proteins (Stano et al., 2016). Another repository, Dr.VIS provide viral integration sites associated with human diseases (Yang et al., 2015; Zhao et al., 2012a). Another, Human Virome Protein Cluster Database (HVPC) for characterization and annotation of the human virome is constructed (Elbehery et al., 2018).

Additionally, different virus and family-specific resources were also developed. Like, flavivirus-specific resources were developed, i.e., FLAVIdB (Olsen et al., 2011), and Flavitrack (Misra and Schein, 2007). Different databases such as DenvInt (Dey and Mukhopadhyay, 2017), DenHunt (Karyala et al., 2016), and Dengue Genographic Viewer (DGV) (Yamashita et al., 2016a) centric to dengue virus was constructed. Further, the Influenza Virus Database (IVDB) platform for genomic and phylogenetics of the Influenza A Virus (Chang et al., 2007) developed. FluGenome, a web-based tool for influenza A virus genotyping was designed (Lu et al., 2007). Also, HCV specific resources such as the Los Alamos HCV Sequence Database (Kuiken et al., 2008; Kuiken et al., 2005), the European hepatitis C virus database (euHCVdb) (Combet et al., 2007) was developed. Further, HFV/Ebola Database, a central repository that provides annotated sequences and analysis tools for Hemorrhagic fever viruses (HFVs). It presents a set of ~80 viral species comprising five different families: *Arena-*, *Bunya-*, *Flavi-*, *Filo-* and *Togaviridae* (Kuiken et al., 2012). CoVDB, a resource for coronavirus genes and genomes was built (Huang et al., 2008). Likewise, An Ebola virus-centered knowledge base provides EBOV genes, protein domains, and genomic information (Kamdar and Dumontier, 2015). Another resource, the HIV database provides data on genetic sequences and immunological epitopes (Kuiken et al., 2003). Moreover, a viral protein domain database (VIP DB) providing protein functions and interaction partners is developed (Chen et al., 2012). Up to now, few papillomavirus-related resources were also developed. The Papillomavirus Episteme (PaVE) is developed that mainly provide papillomavirus genomic and proteomic content (Van Doorslaer et al., 2017; Van Doorslaer et al., 2013). Another, the Human papillomavirus T cell Antigen Database (HPVdb), which hosts antigen and epitope entries was constructed (Zhang et al., 2014).

Apart from the worldwide development of viral resources, there are also efforts from India in the field of viral informatics. Among these, VirGen is an annotated and curated database comprising complete genome sequences of viruses (Kulkarni-Kale et al., 2004). EbolaVCR was constructed, which provide peptide or epitope-based vaccine candidates, and putative siRNAs against the ebola viruses (Dhanda et al., 2016). Another resource, ZikaBase was established, which is a database of the ZIKV-Human interactome map (Gurumayum et al., 2018). Along with this, we have also developed most of the viral computational resources from India. These are mainly focused around different aspects associated with viruses, i.e., like RNAi based; related with siRNAs and miRNAs such as VIRsiRNAdb: a curated database of experimentally validated viral siRNA/shRNA (Thakur et al., 2012c), VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses (Qureshi et al., 2013b), HIVsirDB: A Database of HIV Inhibiting siRNAs (Tyagi et al., 2011), VIRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets (Qureshi et al., 2014a). Further, antiviral peptides-based resources such as AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses (Qureshi et al., 2014d), AVPpred: collection and prediction of highly effective antiviral peptides (Thakur et al., 2012a), HIPdb: A database of experimentally validated HIV inhibiting peptides (Qureshi et al., 2013a). Likewise, antiviral compound-based resources developed are AVCpred: an integrated web server for prediction and design of antiviral compounds (Qureshi et al., 2017), HIVprotI: an integrated web-based platform for prediction and design of HIV proteins inhibitors (Qureshi et al., 2018). Other resources such as MSLVP for the prediction of multiple subcellular localization of viral proteins using a support vector machine (Thakur et al., 2016), vhfRNAi: A web-platform for analysis of host genes involved in viral infections discovered by genome-wide RNAi screen (Thakur et al., 2017). and ViralEpi v1.0: an integrated resource of viral epigenomic methylation profiles from diverse diseases (Khan et al., 2016) was also established. Moreover, we have also developed *in-silico* resources dedicated to the putative therapeutics and epitopes for different infectious and pathogenic viruses. An integrated Zika virus resource (ZikaVR) dedicated to the genomic, proteomic, and therapeutic knowledge (Gupta et al., 2016), NipahVR: a resource for multi-targeted solutions for Nipah virus (Gupta et al., 2020b), Likewise, a computational resource (CoronaVR) and analysis of epitopes and therapeutics for Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) (Gupta et al., 2020a) is developed.

PART-II

Next-generation sequencing (NGS) and metagenomics in virology and virome exploration

Viruses are known to regulate human systems, influence immunity, affect human health, and are linked with distinct diseases (Cadwell, 2015; Paez-Espino et al., 2016; Virgin, 2014; Wylie et al., 2012). Specifically, RNA viruses pose a boundless threat for all (Manso et al., 2017; ME et al., 2013; Wolf et al., 2018). All the viruses (both prokaryotic as well as eukaryotic) found in human altogether make the complete human virome and is an integral part of the host-microbiome (Wylie et al., 2012). Viruses also have a crucial role in impacting microbial community structure mainly due to the bacteriophages (Wylie et al., 2012). However, a large portion of viruses is underestimated and still needed to discover (Shi et al., 2018; Wolf et al., 2018; Woolhouse et al., 2012). Hence, it's very important to explore and investigate the clinical as well as environmental virome (Woolhouse et al., 2012).

High throughput sequencing technologies provided great opportunities to study clinically important viruses such as HIV, influenza virus, HCV, HPV, HBV, etc., in application to diagnostic virology (Barzon et al., 2013; Barzon et al., 2011b; Beerenwinkel et al., 2012; Capobianchi et al., 2013; Quiñones-Mateu et al., 2014; Wylie et al., 2013). This majorly includes virus detection and discovery of novel viral pathogens from clinical specimens, investigating viral genetic diversity (quasispecies), viral genome reconstruction, characterization of virome (viral metagenomics), viral transmission and to monitor the emergence of antiviral drug resistance mutations in diseased conditions (Barzon et al., 2011b; Capobianchi et al., 2013). Also, demonstrated to be useful to detect oncoviruses. Moreover, NGS offers a powerful, ultrasensitive, and essential diagnostic tool with the potential to explicate a complete spectrum of viruses or pathogens that might not be possible with conventional strategies like PCR, microarray, or Sanger sequencing (Barzon et al., 2011b; Beerenwinkel et al., 2012; Capobianchi et al., 2013; Quiñones-Mateu et al., 2014).

For various vital applications, human virome analysis and surveillance system are significantly important (Anderson et al., 2003; Woolhouse et al., 2012; Zou et al., 2016). This will increase the understanding of viral community structure, could be

used to lower the risk of future viral outbreaks, to advance timely disease diagnostics and surveillance and to drive for new viral therapeutics. Moreover, virome characterization is also extremely imperative in transfusion medicine and blood safety (Moustafa et al., 2017; Sauvage and Eloit, 2016; Sauvage et al., 2016). Simultaneously, this could also affect the treatment of various disorders/diseases in patients (Wylie et al., 2012).

In this effort, the Global Virome Project (GVP) (http://www.globalviromeproject.org/) was established for 10 years to hunt and discover unknown viruses across the globe (Carroll et al., 2018). Metagenomic approaches have allowed the discovery of viruses in clinical or environmental samples very rapidly and cheaply. There are various studies which have focused on the exploration of DNA viruses (DNA virome) (Foulongne et al., 2012b; Moustafa et al., 2017; Paez-Espino et al., 2016; Reyes et al., 2010; Wylie et al., 2014). Moreover, the rich diversity of RNA viruses could be explored using the meta-transcriptomic approaches (Manso et al., 2017; Shi et al., 2018; Zou et al., 2016). Studies have shown the metagenomic application to unearth RNA virus genomes (Manso et al., 2017).

Major software tools developed for different sequencing data analysis

Quality control and assessment

Distinct next-generation platforms also suffer from a range of sequencing errors, contaminations, and artifacts. This mainly includes low-quality bases, G/C bias, repeats, homopolymers, duplication, primer/adaptor sequences contaminations, etc. (Bragg et al., 2013; Laehnemann et al., 2016; Meacham et al., 2011; Minoche et al., 2011; Rieber et al., 2013; Robasky et al., 2014; Ross et al., 2013). However, for the better downstream analysis, quality control of raw data is extremely important. To deal with them and pre-process the NGS reads various tools were developed. These are as follows.

One of the most popular tools is FastQC that provides quality check matrices and profiles of data in the form of automatic reports (https://github.com/s-andrews/FastQC). Trimmomatic is designed to handle and trim Illumina sequencing data (Bolger et al., 2014). Likewise, cutadapt removes undesirable primer and adapter sequences, poly-A tails from high-throughput data (https://github.com/marcelm/cutadapt). Quake

program is developed for the correction of sequencing error (Kelley et al., 2010). QC-Chain, a fast quality control tool for raw sequencing data is developed (Zhou et al., 2013). PRINSEQ, a rapid genomic and metagenomic data pre-processing tool is established (Schmieder and Edwards, 2011). NGSQC Toolkit, it's a toolkit for the quality control of 454 and Illumina NGS data (Patel and Jain, 2012).

Further, there are software tools for long-read sequencing data. To work with Nanopore sequencing data a flexible toolkit poretools, which provide format conversion, data exploration, and visualization utilities are developed (Loman and Quinlan, 2014). Another, NanoOK, a tool for quality and error profile analysis (Leggett et al., 2016). NanoPack is a set of tools for pre-processing and visualization of data from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (De Coster et al., 2018). Further, NanoR, an R package to analyze the nanopore sequencing data is developed (Bolognini et al., 2019). Another R package poRe to analyze, organize, and visualize MinION data is constructed (Watson et al., 2015). Likewise, a toolkit HPG Pore is developed to explore and analyze nanopore data (Tarraga et al., 2016).

(Meta)-genomic mapping, assembly, and processing

Along with the emergence and advancements of high throughput sequencing technologies (second and third generation), various mapping or alignment (Fonseca et al., 2012; Hatem et al., 2013), de novo genome assembly tools and software has been evolved rapidly (Earl et al., 2011; Miller et al., 2010; Senol Cali et al., 2019). Some of the vastly utilized are mentioned here.

For the alignment and mapping of raw data, the most exploited algorithmic approaches were Burrows-Wheeler transform (BWT) and spaced seed (Trapnell and Salzberg, 2009). Exceedingly used tools are BWA or BWA-MEM (Li and Durbin, 2010), Bowtie or Bowtie2 (Langmead and Salzberg, 2012), Minimap2 (Li, 2016), BLAST (Altschul et al., 1990; Camacho et al., 2009), Usearch/Ublast (Edgar, 2010), RAPSearch2 (Zhao et al., 2012c), GraphMap (Sović et al., 2016), BLASR (Chaisson and Tesler, 2012), marginAlign (Jain et al., 2015), etc.

For the assembly of a sequencing data array of assemblers were developed to date. These assemblers are mainly based on two approaches, i.e., Overlap, Layout, Consensus (OLC), and *de Bruijn* graph (dBg) with their hybrids. Among these, highly

recognized and employed short read assemblers are Velvet (Zerbino and Birney, 2008), SOAPdenovo (Li et al., 2010b), Edena (Hernandez et al., 2008), ABySS (Simpson et al., 2009), Iterative De Bruijn graph Assembler (IDBA) (Peng et al., 2012), ALLPATHS (Butler et al., 2008; Maccallum et al., 2009), SPAdes (Bankevich et al., 2012), Minimus (Sommer et al., 2007), SSAKE (Warren et al., 2007), and so on. Similarly, there are distinct long-read assembly tools for the Nanopore and PacBio sequencing data that were also built (Jayakumar and Sakakibara, 2019; Lu et al., 2016). Some of them are Canu (Koren et al., 2017), Racon (Vaser et al., 2017), Miniasm (Li, 2016), Flye (Kolmogorov et al., 2019), HINGE (Kamath et al., 2017), etc. Detailed list is given in **Table 2**.

Table 2. List of long-read assembly tools

Name	***Platform (Nanopore/PacBio)***	***Journal***	***References***
Canu	Both	Genome Res.	(Koren et al., 2017)
Wtdbg2	Both	Nat. Methods	(Ruan and Li, 2020)
Flye	Both	Nat Biotechnol.	(Kolmogorov et al., 2019)
Kermit	Both	Algorithms Mol Biol.	(Walve et al., 2019)
Unicycler	Both	PLoS Comput Biol	(Wick et al., 2017)
Racon	Both	Genome Res.	(Vaser et al., 2017)
Miniasm	Both	Bioinformatics	(Li, 2016)
MHAP	PacBio	Nat Biotechnol.	(Berlin et al., 2015)
HGAP	PacBio	Nat. Methods	(Chin et al., 2013)
FALCON	PacBio	Nat. Methods	(Chin et al., 2016)
HINGE	PacBio	Genome Res.	(Kamath et al., 2017)
ABruijn assembler	Both	PNAS	(Lin et al., 2016)

Additionally, distinct software packages were also developed for the assembly of metagenomic sequencing data from different sequencing platforms (Ayling et al., 2020; Vollmers et al., 2017). Few widely utilized tools are MetaVelvet (Namiki et al., 2012), MEGAHIT (Li et al., 2015)(Li et al., 2016a), metaSPAdes (Nurk et al., 2017), Meta-IDBA (Peng et al., 2011), IDBA-UD (Peng et al., 2012) etc. List of metagenomic assemblers are provided in **Table 3**.

Table 3. List of available metagenomic assemblers

Name	*Algorithm*	*References*
MetaVelvet	dBg (single kmer)	(Namiki et al., 2012)
MetaVelvet-SL (Supervised Learning)	dBg (single kmer)	(Afiahayati et al., 2015)
MEGAHIT	succinct dBg (multiple kmer)	(Li et al., 2015) (Li et al., 2016a)
MegaGTA	succinct dBg	(Li et al., 2017)
SPAdes and metaSPAdes	dBg (multiple kmer)	(Bankevich et al., 2012) (Nurk et al., 2017)
Meta-IDBA	dBg (multiple kmer)	(Peng et al., 2011)
IDBA-UD	dBg (multiple kmer)	(Peng et al., 2012)
Genovo	OLC	(Laserson et al., 2011) (Afiahayati et al., 2013)
MAP	OLC	(Lai et al., 2012)
Omega	OLC	(Haider et al., 2014)
Ray Meta	dBg (single kmer)	(Boisvert et al., 2012)
Snowball	OLC (Iterative joining)	(Gregor et al., 2016)
Xander	dBg+hidden Markov model (HMM)	(Wang et al., 2015a)
PRICE	Hybrid	(Ruby et al., 2013)
MetAMOS	Hybrid Pipeline	(Treangen et al., 2013)
IMP	Hybrid Pipeline	(Narayanasamy et al., 2016)
InteMAP	Hybrid Pipeline (ABySS, IDBA-UD, CABOG)	(Lai et al., 2015)
MetaCRAM	Hybrid	(Kim et al., 2016)

dBg, *de Bruijn* graphs; OLC, Overlap layout consensus

Further, data processing, conversion, and assembly assessment tools were also established. The routinely applied are SAMtools (Li et al., 2009), Bamtools (Barnett et al., 2011), Picard (https://sourceforge.net/projects/picard/files/picard-tools/), BEDTools (Quinlan and Hall, 2010). For the evaluation and comparison of assemblies QUAST (Gurevich et al., 2013) and QUAST-LG (for large genomic assemblies) (Mikheenko et al., 2018) are also developed. These provide a number of assembly matrices like N50, NA50, contig accuracy, coverage, predicted genes, mismatches, etc. Also, for the assessment of metagenomic assemblies MetaQUAST is

developed (Mikheenko et al., 2016). For visualization, some specific tools were also developed to depict and plot a large amounts of data like Circos (Krzywinski et al., 2009), Graphlan (Asnicar et al., 2015), krona (Ondov et al., 2011), etc.

Virus and phage specific NGS tools and software

To date, there are also tools or pipelines available for the viral NGS data analysis (**Table 4**) (Nooij et al., 2018; Orton et al., 2016). In last some years, certain efforts are made for the development of viral NGS data assembly tools, i.e., VICUNA (Yang et al., 2012), Arapan-S (Sahli and Shibuya, 2012), VGA (Mangul et al., 2014), IVA (Hunt et al., 2015), VirAmp (Wan et al., 2015), and V-GAP (Nakamura et al., 2016), each having evident advantages along with distinct limitations like VICUNA works only with non-repetitive genomes, Arapan-S mainly deals with long reads, IVA works with RNA virus genomes and so on.

Likewise, for detection of viruses, assorted algorithms were developed namely, VirusHunter (Zhao et al., 2013), for identification of novel viruses using long-read next-generation sequencing platform data; VirusFinder (Wang et al., 2013), software for detection of viruses and integration sites; and VirFind (Ho and Tzanetakis, 2014), a bioinformatics tool specifically for virus detection and discovery. Also, for viral variant detection some methods were developed, i.e., ViVaMBC (Verbist et al., 2015a), a virus variant model-based clustering method for identifying and quantifying viral variants at the codon level; VirVarSeq (Verbist et al., 2015c), a low-frequency virus variant detection pipeline and ViVan (Isakov et al., 2015), a pipeline facilitating the identification, characterization, and comparison of sequence variance in deep sequenced virus populations.

Similarly, for metagenomic and virome studies certain algorithms were developed, i.e., VirusTAP (Yamashita et al., 2016b), which deals with viral genome-targeted assembly; VIP (Li et al., 2016c), an integrated pipeline for metagenomics of virus identification and discovery; ViromeScan (Rampelli et al., 2016), to explore and taxonomically characterize the virome from metagenomic reads; Metavir (Roux et al., 2011) and Metavir 2 (Roux et al., 2014), for viral sequence analysis, taxonomic profiling and assembled virome analysis. Furthermore, a generalized pathogen identification cloud compatible bioinformatics pipeline was also developed named SURPI ("sequence-based ultra-rapid pathogen identification") (Naccache et al., 2014). The exhaustive list of currently available viruses specific NGS tools is provided in **Table 4**.

Table 4. List of viruses specific NGS and metagenomic data analysis tools

Name	*Application*	*Reference*
ViVan	Identification, characterization and comparison of sequence variance in deep sequenced virus populations	(Isakov et al., 2015)
VirVarSeq (Q-cpileup)	A low-frequency virus variant detection pipeline (Quasispecies)	(Verbist et al., 2015c)
ViVaMBC	Estimating viral sequence variation in complex populations using model-based clustering	(Verbist et al., 2015a)
VirFind	Virus detection and discovery pipeline	(Ho and Tzanetakis, 2014)
Virus Hunter	Identification of novel viruses (Roche/454)	(Zhao et al., 2013)
VICUNA	Consensus assembly of ultra-deep sequence derived from diverse viral populations	(Yang et al., 2012)
V-GAP	Pipeline to assemble small viral genomes with good reliability using a resampling method from shotgun data	(Nakamura et al., 2016)
V-Phaser and V-Phaser 2	Highly sensitive and specific detection of rare variants in mixed viral populations	(Macalalad et al., 2012) (Yang et al., 2013)
BATVI	Fast, sensitive and accurate detection of virus integrations	(Tennakoon and Sung, 2017)
VERSE	A novel approach to detect virus integration in host genomes	(Wang et al., 2015b)
VirusFinder	For efficient and accurate detection of viruses and their integration sites in host genomes	(Wang et al., 2013)
Metavir and Metavir 2	A web server dedicated to the analysis of viral metagenomes (viromes), New tools for viral metagenome comparison and assembled virome analysis	(Roux et al., 2011) (Roux et al., 2014)

Name	*Application*	*Reference*
ViromeScan	Tool for metagenomic viral community profiling (eukaryotic viruses)	(Rampelli et al., 2016)
ViraPipe	Scalable parallel pipeline for viral metagenome analysis (distributed Spark computing cluster)	(Maarala et al., 2018)
MG-Digger	An Automated Pipeline to Search for Giant Virus-Related Sequences in Metagenomes	(Verneau et al., 2016)
viGEN	An Open Source Pipeline for the Detection and Quantification of Viral RNA in Human Tumors	(Bhuvaneshwar et al., 2018)
Vy-PER	Eliminating false positive detection of virus integration events (virus/host chimera detection)	(Forster et al., 2015)
HGT-ID	An efficient and sensitive workflow to detect human-viral insertion sites	(Baheti et al., 2018)
Virus-Clip	Fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability	(Ho et al., 2015)
QuRe	Software for viral quasispecies reconstruction	(Prosperi and Salemi, 2012)
QuasQ	Viral quasispecies inference from 454 pyrosequencing	(Poh et al., 2013)
Virana	Sensitive detection of viral transcripts in human tumor transcriptomes	(Schelhorn et al., 2013)
ViralFusionSeq	Accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution	(Li et al., 2013a)
viral profile HMMs ("vFams")	Profile hidden Markov models for the detection of viruses within metagenomic sequence data	(Skewes-Cox et al., 2014)

Name	*Application*	*Reference*
Taxonomer	An interactive metagenomics analysis portal for universal pathogen detection and host mrna expression profiling	(Flygare et al., 2016)
ProViDE	A software tool for accurate estimation of viral diversity in metagenomic samples	(Ghosh et al., 2011)
VirusSeq	Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue	(Chen et al., 2013)
HIVID	Efficient method to detect HBV integration using low coverage sequencing	(Li et al., 2013b)
VirAmp	A galaxy-based viral genome assembly pipeline	(Wan et al., 2015)
IRMA pipeline	Solves the problem of viral variation by the iterative optimization of read gathering and assembly, IRMA also focuses on quality control, error correction, indel reporting, variant calling and variant phasing	(Shepard et al., 2016)
IVA	Accurate de novo assembly of RNA virus genomes	(Hunt et al., 2015)
VGA	Accurate viral population assembly from ultra-deep sequencing data, uses an expectation-maximization algorithm to estimate abundances of the assembled viral variants in the population	(Mangul et al., 2014)
drVM	Tool for efficient genome assembly of known eukaryotic viruses from metagenomes	(Lin and Liao, 2017)
Arapan-S	Fast and highly accurate whole-genome assembly software for viruses and small genomes	(Sahli and Shibuya, 2012)
SAVAGE	A computational tool for reconstructing individual haplotypes of intra-host virus strains without the need for a high-quality reference genome (De novo assembly of viral quasispecies)	(Baaijens et al., 2017)

Name	*Application*	*Reference*
TAR-VIR	A pipeline for targeted viral strain reconstruction from metagenomic data (optimized for identifying RNA viruses from metagenomic data)	(Chen et al., 2019)
VirusTAP	A web-based integrated NGS analysis tool for the viral genome (virus genome-targeted assembly pipeline)	(Yamashita et al., 2016b)
virMine	Automated detection of viral sequences from complex metagenomic samples	(Garretto et al., 2019)
VirMAP	Maximal viral information recovery from sequence data, merge nucleotide and protein information to taxonomically classify viral reconstructions	(Ajami et al., 2018)
VIROME	A standard operating procedure for analysis of viral metagenome sequences	(Wommack et al., 2012)
VMGAP	An automated tool for the functional annotation of viral Metagenomic shotgun sequencing data	(Lorenzi et al., 2011)
VIP	An integrated pipeline for metagenomics of virus identification and discovery	(Li et al., 2016c)
VirSorter	Mining viral signal from microbial genomic data	(Roux et al., 2015)
VirFinder	A novel k-mer based tool for identifying viral sequences from assembled metagenomic data	(Ren et al., 2017)
VirusSeeker	A computational pipeline for virus discovery and virome composition analysis	(Zhao et al., 2017)
VirusDetect	An automated pipeline for efficient virus discovery using deep sequencing of small rnas	(Zheng et al., 2017)
iVirus	Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure	(Bolduc et al., 2017)
FastViromeExplorer	A pipeline for virus and phage identification and abundance profiling	(Tithi et al., 2018)

Name	*Application*	*Reference*
Vipie	Web pipeline for parallel characterization of viral populations from multiple NGS samples	(Lin et al., 2017)
VirGenA	A reference-based assembler for variable viral genomes, can separate mixtures of strains of different intraspecies genetic groups	(Fedonin et al., 2019)
MetaPORE	Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis, web-based pipeline for real-time bioinformatics analysis on a computational server or laptop	(Greninger et al., 2015)
Clinical PathoScope	Rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data	(Byrd et al., 2014)
CaPSID	A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes,	(Borozan et al., 2012)
SURPI	A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples	(Naccache et al., 2014)
MiCoP	Microbial community profiling method for detecting viral and fungal organisms in metagenomic samples	(LaPierre et al., 2019)
MetaPhinder	Identifying Bacteriophage Sequences in Metagenomic Data Sets.	(Jurtz et al., 2016)
PhageWeb	Web server for identification of prophages	(de Sousa et al., 2018)
Phage_Finder	Identification of prophage regions in bacterial genome	(Fouts, 2006)
Phage Hunters	Computational strategies for finding phages in large-scale ‘omics datasets	(Hurwitz et al., 2018)
MARVEL	Prediction of bacteriophage sequences in metagenomic bins	(Amgarten et al., 2018)

Name	*Application*	*Reference*
Prophage Hunter	An integrative hunting tool for active prophages	(Song et al., 2019)
Prophage Finder	A prophage loci prediction tool	(Bose and Barber, 2006)
Prophinder	Tool for prophage prediction	(Lima-Mendez et al., 2008)
PHAST	A fast phage search tool	(Zhou et al., 2011)
PhiSpy	A novel algorithm for finding prophages in bacterial genomes	(Akhter et al., 2012)
PHASTER	A better, faster version of the PHAST phage search tool	(Arndt et al., 2016)
PHANOTATE	A novel approach to gene identification in phage genomes	(McNair et al., 2019)
HPViewer	Sensitive and specific genotyping of human papillomavirus in metagenomic DNA	(Hao et al., 2018)
HPVDetector	NGS-based approach to determine the presence of HPV and their sites of integration	(Chandrani et al., 2015)

Rationale

HPVs are the oncoviruses responsible and associated with the diverse carcinomas namely cervical, HNSCC, penile, anal, vulvar, etc. Persistence infection of these viruses is critical and decisive in the progression of cancer (Crosbie et al., 2013; zur Hausen, 2002, 2009). HPV infection is very common in women. However, only a few (~1% cases) use to progress towards high-grade carcinoma, i.e., CIN2 and CIN3 (Schiffman et al., 2011). There are events associated with HPV infection and progression, which could be used as valuable biomarkers (Schiffman and Wentzensen, 2013). Likewise, three VLP-based prophylactic vaccines were also developed and approved to prevent HPVs infection. However, these vaccines are not able to eradicate the lesions and not-effective in abolishing established infections (Chabeda et al., 2018; Cheng et al., 2018; Dadar et al., 2018; Hildesheim et al., 2007; Hu and Ma, 2018). Thus, there is still a requirement for effective therapeutics and drugs to combat HPV infection and associated carcinomas. Despite substantial efforts to eradicate the HPVs, there is no resource available for the therapeutically imperative biomarkers. Moreover, comprehensive computational analysis of HPV-associated events is also lacking. Further, there is no existing resource for the potential HPV therapeutics and vaccine epitope candidates are available.

Viruses are the most abundant and widely distributed biological bodies on earth. However, the large portion is still unknown and undiscovered. NGS and metagenomics technologies could be utilized in diagnostics and to explore the diversity of viruses from distinct ecosystems. To analyze the NGS data, numerous computational tools were also developed to date. Along with the development of diverse tools, various studies also evaluated the comparative performance of distinct algorithms (Bao et al., 2011; Earl et al., 2011; Finotello et al., 2012; Magoc et al., 2013; Salzberg et al., 2012; Zhang et al., 2011). However, these studies mostly used human, bacterial, and plant NGS data (Barthelson et al., 2011; Earl et al., 2011; Salzberg et al., 2012). Moreover, there is a lacuna of reasonable assessment studies for the comparison of existing assembly algorithms on viral NGS data. Therefore, it is important to analyze distinct assemblers on viral sequencing data from different platforms. Furthermore, some of the viral metagenomic data analysis pipelines were also developed. Nevertheless, distinct challenges and gaps were still existing (Lambert et al., 2018; Rose et al., 2016). Like, some works with either environmental

or clinical data, need computing proficiency, lack of quality control steps, host subtraction step, etc. (Lambert et al., 2018; Rose et al., 2016). Based on the above-mentioned gaps and scope, we have framed our aims and objectives.

Aims and Objectives

The broad title of the research work is "***Development of integrative bioinformatics resources for the analysis of viral next generation sequencing (NGS) data and human papillomaviruses (HPVs)*****"**.

The following is the list of objectives set for the completion of the proposed research work:

- Development of Human Papillomaviruses (HPVs) mediated disease biomarker knowledgebase
- Systematic meta-analysis of Human genes disrupted due to HPVs associated events
- Development of HPVs genomic and therapeutic resource
- Benchmarking of *de novo* genome assemblers for the viral next generation sequencing (NGS) data
- Development of bioinformatics tool or pipeline for viral NGS data analysis: Implication in HPV research

Development of HPV mediated disease biomarker knowledgebase

Chapter 2. Development of HPV mediated disease biomarker knowledgebase

Introduction

Human papillomaviruses (HPVs) are the double-stranded DNA (dsDNA) circular genome virus from the *Papillomaviridae* family. HPV genomes are 8 kb in length that consists of 8-10 open reading frames (ORFs) along with one non-coding regulatory long control region (LCR). Two coding regions (early (E) and late (L)) mainly encode eight well-defined proteins i.e. E1, E2, E4, E5, E6, E7, L1, and L2. Among these, E1, E2, E4 are mainly involved in functional regulation, E5, E6, E7 are known oncogenes, and L1, L2 are two viral capsid genes (de Villiers et al., 2004; Doorbar, 2006; Munoz et al., 2003).

HPVs are known to cause various carcinomas however primarily prevalent in cervical oncogenesis. These mainly utilize host machinery for functioning and survival. In HPV mediated oncogenesis, the expression and function of two oncoproteins E6 and E7 are most critical in the tumor progression. These proteins are used to accelerate proliferation, immortalization, malignancy, and target different cellular components. They principally obstruct two important tumor suppressor proteins i.e. p53 and retinoblastoma (Rb) family. This inhibits proteolytic degradation, cell cycle arrest, abrogates apoptosis, etc. (Mantovani and Banks, 2001; Moody and Laimins, 2010; Munger et al., 2001; Vande Pol and Klingelhutz, 2013). Cervical cancer (CaCx) is one of the most common reasons of mortality among women all over the world. HPV infection and persistency play a major role in invasive carcinoma (Crosbie et al., 2013; Doorbar, 2006; Doorbar et al., 2012; Schiffman and Wentzensen, 2013). These well-known factors also advanced the prevention and screening strategies against HPV associated cancers (Boulet et al., 2008; Sahasrabuddhe et al., 2011). Nevertheless, this still requires further attention and remains the prevailing cause of cancer deaths. Furthermore, there is also a need to have an operative way to differentiate between transient infection, pre-cancer, and eventually high-grade carcinoma. Moreover, there is no effective treatment available to eradicate cancer (Schiffman and Wentzensen, 2013; Woodman et al., 2007).

To this end, there are HPV related events and elements that can discriminate neoplastic progression and could act as potential biomarkers (Schiffman and Wentzensen, 2013). This mainly includes viral DNA integration, viral methylation, and abrupt expression of cellular miRNAs (Sahasrabuddhe et al., 2011; Schiffman and Wentzensen, 2013). Integration is one of the most crucial events pertinent to HPV mediated carcinoma (Akagi et al., 2014; Klaes et al., 1999; Parfenov et al., 2014; Schmitz et al., 2012; Thorland et al., 2003; zur Hausen, 2002). HPVs are known to integrate into the host genome during the tumor progression and stabilize and enhance transcription of HPV oncogenes (E6 and E7) (Jeon and Lambert, 1995; Wentzensen et al., 2004). In turn, disturb cellular genomic instability (Akagi et al., 2014; Thorland et al., 2003). Likewise, methylation (viral or host) also modulates HPV transcriptional regulation (Brandsma et al., 2009; Clarke et al., 2012; Turan et al., 2007). Further, HPV oncogenes are known to interact with distinct cellular targets (Doorbar, 2006) and may affect tumor-suppressive or oncogenic miRNAs (oncomirs) regulation (Gomez-Gomez et al., 2013; Reshmi and Pillai, 2008; Zheng and Wang, 2011). These miRNAs could alter and be involved in various molecular mechanisms like cell cycles, growth, apoptosis, cell proliferation, signalling, etc. (Lee et al., 2008a; Winter and Diederichs, 2011).

Numerous studies advocate the importance of these factors. However, there is no such computational resource available specific to these therapeutically important alternative biomarkers. Moreover, the overall picture related to these events in different epidemiological conditions is largely unexplored. To overcome this paucity, multi-targeted web-based platform with a unique focus on different biomarkers and analysis is developed to facilitate the further research.

Materials and Method

Biomarker data collection and curation

HPV-associated literature specific to particular events were systematically searched on the PubMed repository maintained by the National Center for Biotechnology Information (NCBI), a division of the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). Exhaustive quest utilizing different keywords for all three components namely (1) HPVs integrations (2) HPVs methylation patterns and (3) abnormal expression of host miRNAs due to HPV infection.

HPV integrations and breakpoints

Published literature was precisely searched via query with a set of keywords "((((((HPVs) OR human papillomaviruses) OR human papillomavirus) OR HPV*)) AND (((cancer) OR carcinoma))) AND ((integration*) OR breakpoint*)". In total, 755 scientific articles were retrieved from which review articles (117) were excluded. Finally, 638 research papers were examined for the retrieval of relevant data and meta-information. Further, studies that only provide the status (presence or absence) of HPV genes but not exact integration sites and coordinates were also excluded. Comprehensive information and clinical details after a careful reading of literature were extracted like HPV genotypes (i.e. HPV16, 18), HPV regions (e.g., E6, E7), viral integration sites and breakpoints (e.g., 450:474), human chromosome (e.g., 8, 3) and coordinates (e.g., 26257343:26257366), cytobands (e.g., q23), target region or genes (e.g., RAD51B, MYC), linked fragile sites (e.g., FRA8C, FRA8D), detection approach (e.g., RT-PCR, APOT assay, RNA-seq), cancer types or histology (e.g., cervical cancer), sample type or specimen (e.g., tumor biopsy, HeLa cells), etc. Complete data was later cross-checked and curated to rectify inconsistencies and remove any errors.

HPVs DNA methylation

Correspondingly, studies related to HPVs methylation were searched utilizing a blend of different words. The query is "(((((((HPVs) OR human papillomaviruses) OR human papillomavirus) OR HPV*))) AND (((cancer) OR carcinoma))) AND methylation". Overall, 289 research articles excluding 31 reviews were screened. Further, many literatures were not considered as they only provide methylation of host DNA, which is not covered in the current study. Detail data of HPV methylation that includes HPVs types (i.e. HPV18), HPV gene (e.g. E2, E6), methylation pattern (e.g. Hypo methylation), methylation detection method (e.g., bisulfite sequencing), specimen type (e.g., clinical biopsy), related carcinoma or grade (like cervical cancer or cervical intraepithelial neoplasia (CIN)), etc. is extracted, curated and provided.

Host miRNAs regulations

Similarly, to find all articles from PubMed related to aberrant expression of host miRNAs due to HPV infection is searched via a combination of keywords "((((((HPVs) OR human papillomaviruses) OR human papillomavirus) OR HPV*))

AND ((cancer) OR carcinoma)) AND ((((microRNA) OR miRNA) OR microRNAs) OR miRNAs)". Inclusively, 123 articles including 24 reviews were retrieved. We have collected information that comprehends miRNAs, regulation patterns (expression), miRBase id, cellular location and coordinates, and linked carcinoma. Added, external resources were also linked and integrated. Moreover, target genes of diverse miRNAs were obtained and explored utilizing MiRTarBase (Hsu et al., 2014).

Web-interface

The back-end of the HPVbase web interface is supported through the open source LAMP (Linux-Apache-MySQL-PHP) solution stack. Further, front-end is developed using web and scripting languages i.e. HTML, javascript, PHP, and Perl. Additionally, a lightweight browser is constructed to represent the specific biomarkers and descriptive information interactively utilizing JavaScript Object Notation (JSON) data format employing JBrowse (Skinner et al., 2009). Data files were converted into a gene feature file (GFF3) format using Perl script to use in JBrowse. The whole system is accommodated on the IBM machine with the Red Hat Enterprise Linux 5 environment having Apache 2.2.17 server, MySQL (5.0.51 b) and PHP (5.2.14).

Results and Discussion

HPVbase architecture

It is a web-based resource for the potential biomarkers (viral and cellular) associated with the pathogenicity of HPV-linked carcinoma. The resource is organized into the three distinct sections for all three components namely integration events, HPV methylations, and aberrant expression of host miRNAs (**Figure 4**). Later, these sections are categorized into different subsections specific to HPV types and carcinoma. Data from each biomarker are represented in tabular as well as interactive browser. Further, HPVbase also provides browsing, searching and sorting facility for easy data retrieval. Users can explore, and search data using related keywords i.e. particular HPV types (like HPV16), genomic region (E6, E1), cytoband (8q24.21), detection method (RNA-seq), target genes (TP63) etc. For these two user-friendly search tools namely integration search and advance search is implemented at web server. In first, integration site information can be explored based on different keywords utilizing exact or containing mode through restricting to a diverse number of fields. In an advanced search tool, users can perform search employing logical

operators (AND/OR). This allows generating queries via combining specific keywords to filter out the search. The complete resource is freely available at http://crdd.osdd.net/servers/hpvbase.

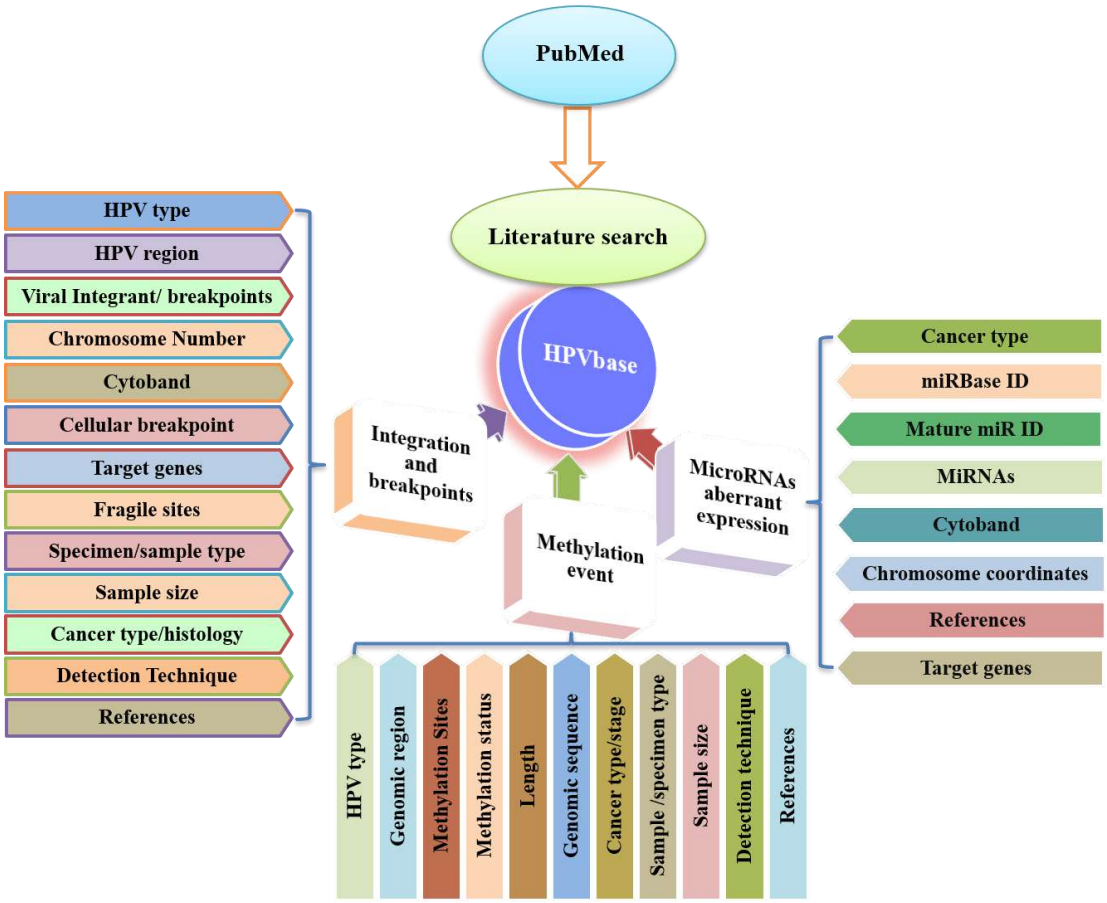


Figure 4. Layout depicting overall organization of HPVbase

HPV integration sites

Various studies have advocated the significant role of HPV integration into the host genome in the advancement of various carcinomas. However, the complete mechanism of integrations is unclear. We have designed an inclusive resource and browser that contains HPV integrates. Overall, 1257 integration sites and breakpoints linked to the different conditions were collected. Among all, most of the integration data mainly belong to the HPV16 and HPV18. Varied information like viral integrate, cellular location, target genomic region, linked fragile site, specimen or sample type, sample size, cancer types, detection approach used, etc. are provided.

An interactive browser is constructed to find and browse HPV integration events. All corresponding HPV integration information is provided through various color-encoded blocks. This is depicted in a different set of tracks (**Figure 5**). It can also be utilized to compare and map experimental integration data. Moreover, integration data can also be explored along with additional annotations and external links in user-affable tabular format (**Figure 6**). Distribution and abundance of integration sites pertaining to different HPV genotypes are shown in **Figure 7**. These are mainly belonging to HPV16 (954), 18 (216), 33 (33), and 45 (33). Further, chromosome loci and disease-specific occurrence of integration events were analyzed and stated (**Figure 8-9**). We specify that some loci regions mainly 8q24.21, 3q28, 13q22.1, 9q22.33, and 14q24.1 are the preferential target and have higher viral integration in different carcinomas (**Figure 9**). Interestingly, all highly preferred loci belong to the q arm of chromosomes. Genome-wide HPV16 integration is also depicted using Circos with cytobands and disrupted genes (**Figure 10**).

Furthermore, integration frequency on the entire host genome is also imperative. For this, HPV type-wise distribution of integration sites on distinct human chromosomes was also analyzed. Chromosomal distribution of integration sites of HPV16, 18, 45, and 33 are illustrated in **Figure 11**. We have identified the most liable target regions for genomic instability from the human genome. Although these sites are present and covering the entire human genome, some of the regions are displaying a higher tendency towards disruption. These prominent regions can be considered as hot-spot for cancer research. HPV16 primarily prefer integration on chromosome 3, 9, 2, 1, and 8. Likewise, HPV18 used to favor 8^{th}, 1^{st}, 2^{nd}, 5^{th} and 3^{rd} chromosomes.

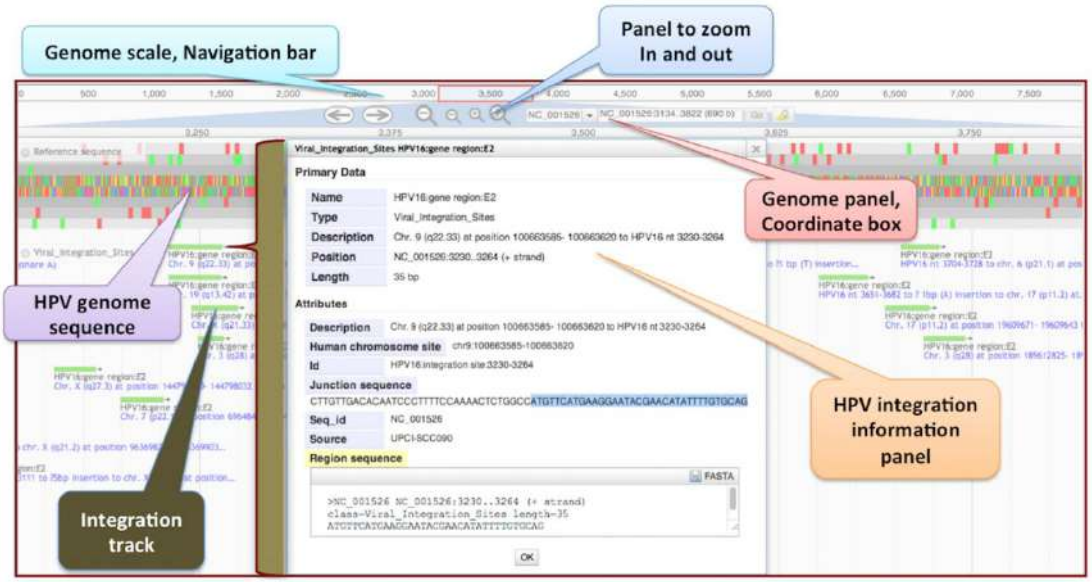


Figure 5. Screenshot showing integration site mapped on the HPV16 (NC_001526.2) reference genome, along with detailed description such as human genome position, region of HPV genome, junction sequence, source, integrated HPV DNA sequence and length

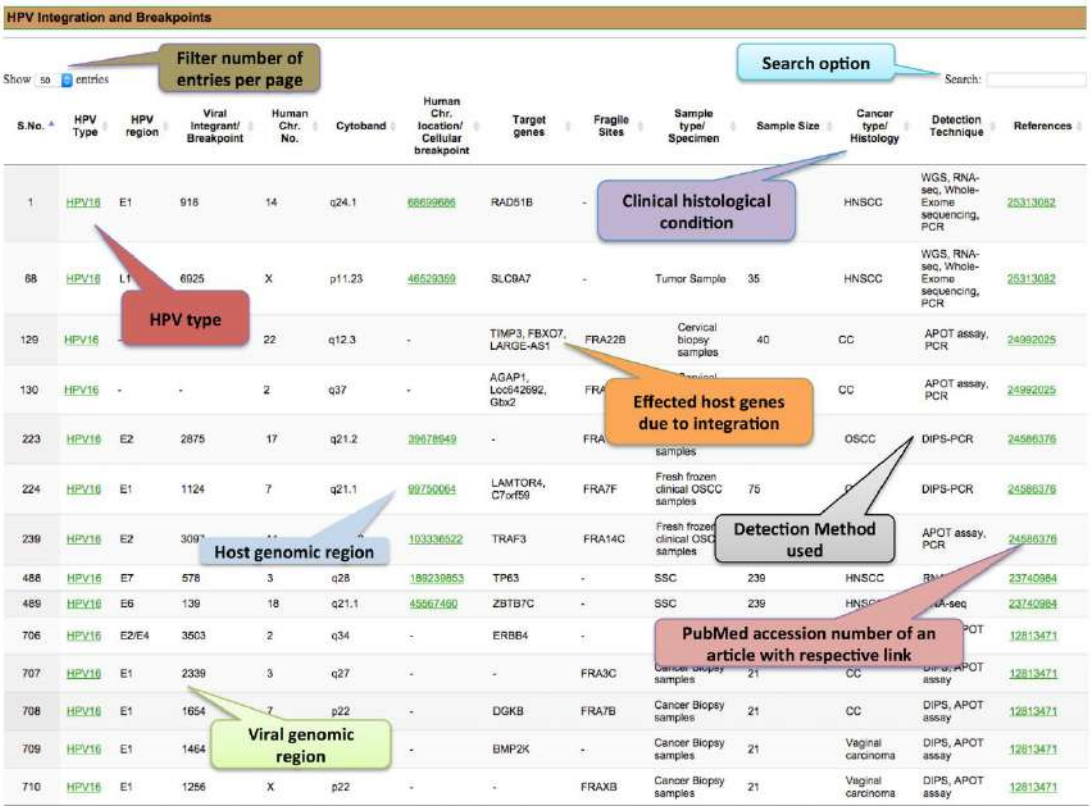


Figure 6. A screenshot depicting integration data with corresponding clinical annotations and reference information in tabular format

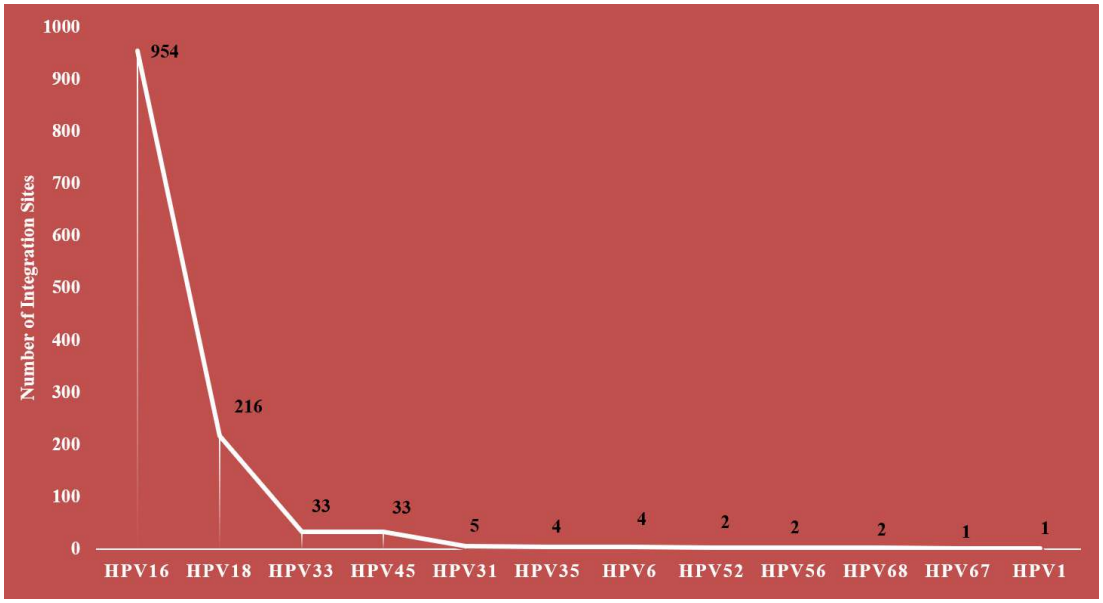


Figure 7. Distribution of integration sites among distinct HPV types

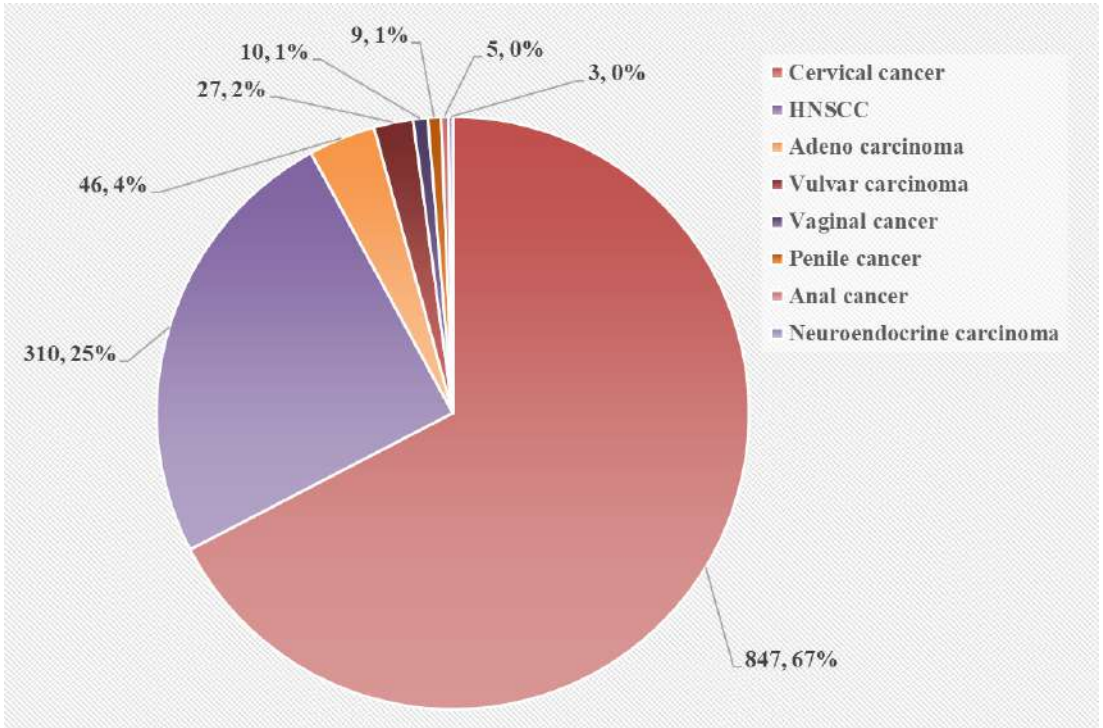


Figure 8. Distribution of integration sites among distinct cancer types

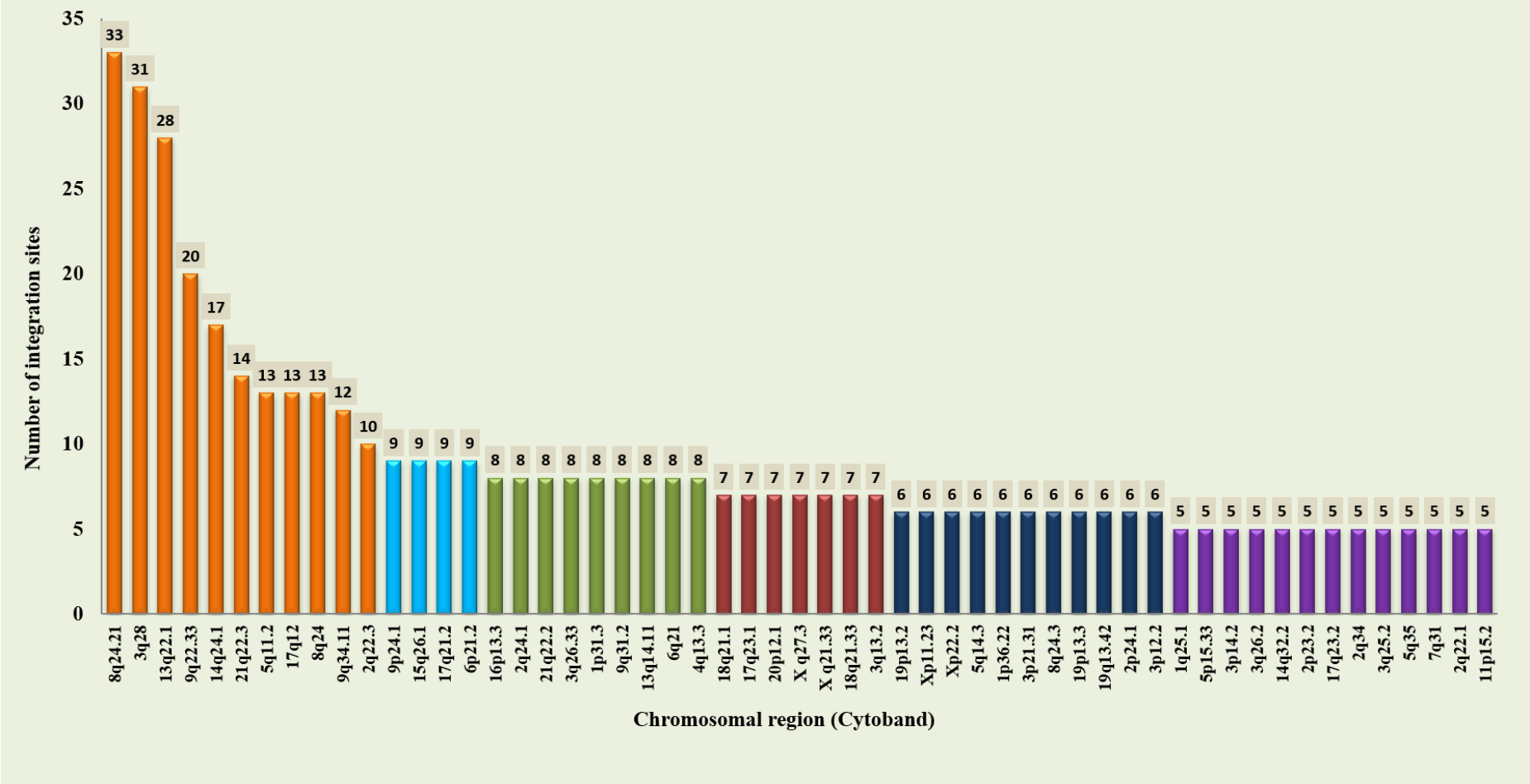


Figure 9. Representing number of integration sites distributed at major genomic loci regions

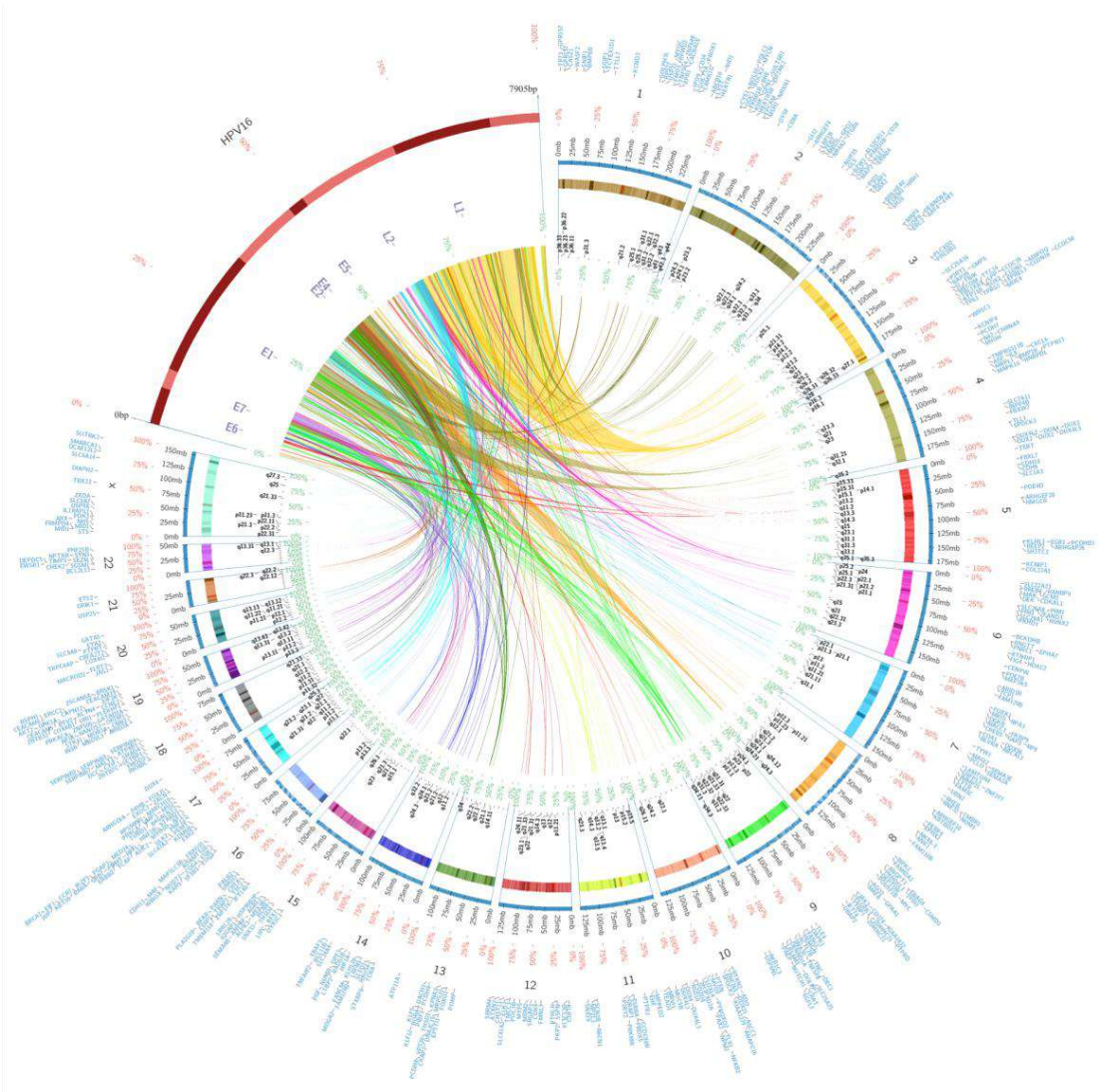
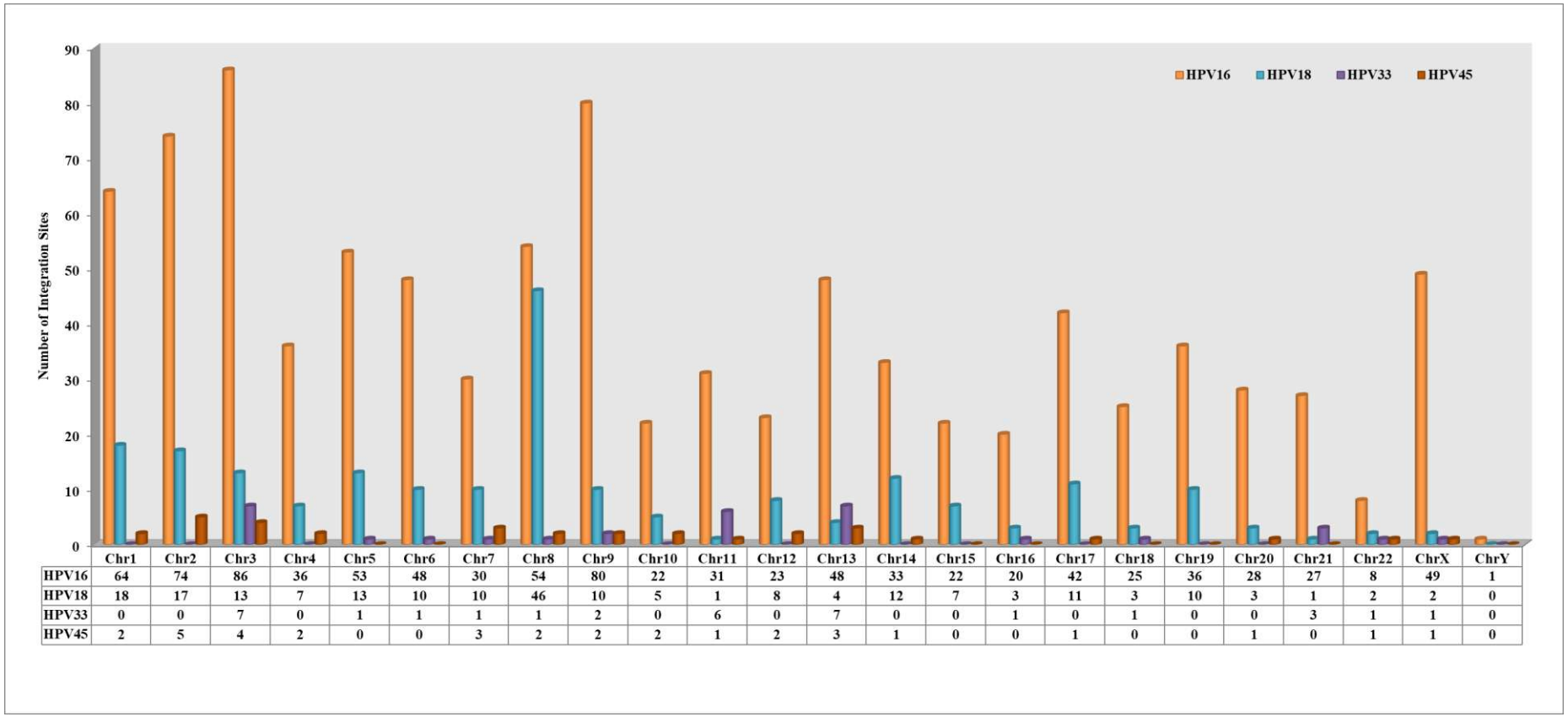


Figure 10. **Circos plot representing genome-wide integration pattern of HPV16 into human genome with chromosomal cytobands and disrupted gene information**. From inside to out, the innermost ring with rainbow color lines (specific for each chromosomes) linking HPV16 genomic coordinates to human idiogram delineating genomic integration sites (chromosome numbers in a clockwise direction, and small red segment within each chromosome indicating the centromere). Dark bands within chromosomes depicting integration hot spots. The second ring represents cytoband information. At last, outermost circle shows the host genes (in blue color) disrupted due to viral integration

Correspondingly, disrupted genes due to HPVs integrations were also described. Firstly, unavailable gene names are extracted from the UCSC genome browser using human genomic coordinates. Subsequently, from total cellular target compendium duplicates, pseudogenes, and (long) non-coding RNAs were removed. Disrupted genes that occur at least twice are depicted in **Figure 12**. Out of these, several genes are related to tumor development as well as regulatory processes, which could be involved cooperatively in the induction of viral oncogenesis. Such as MYC, TP63, RAD51B, FHIT, ETS2, etc. Like, MYC is a viral oncogene homolog play role in apoptosis, cell cycle, and cellular transformation. Tumor protein 63 (TP63) is a member of the transcription factors family mainly regulates neoplastic progression and proliferation. RAD51 paralog B (RAID51B) is a vital element of the DNA repair mechanism and concomitant with the cell apoptosis and cell cycle delay. Likewise, ETS2 and FHIT are tumor-suppressive genes that also involve in the regulation of telomerase and vulnerable to translocations, respectively. Thus, integration events at these regions predominantly signifying their role and association with the different cellular machineries.



	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12	Chr13	Chr14	Chr15	Chr16	Chr17	Chr18	Chr19	Chr20	Chr21	Chr22	ChrX	ChrY
HPV16	64	74	86	36	53	48	30	54	80	22	31	23	48	33	22	20	42	25	36	28	27	8	49	1
HPV18	18	17	13	7	13	10	10	46	10	5	1	8	4	12	7	3	11	3	10	3	1	2	2	0
HPV33	0	0	7	0	1	1	1	1	2	0	6	0	7	0	0	1	0	1	0	0	3	1	1	0
HPV45	2	5	4	2	0	0	3	2	2	2	1	2	3	1	0	0	1	0	0	1	0	1	1	0

Figure 11. Bar chart representing distribution of HPV16, HPV18, HPV33 and HPV45 integration sites on human genome

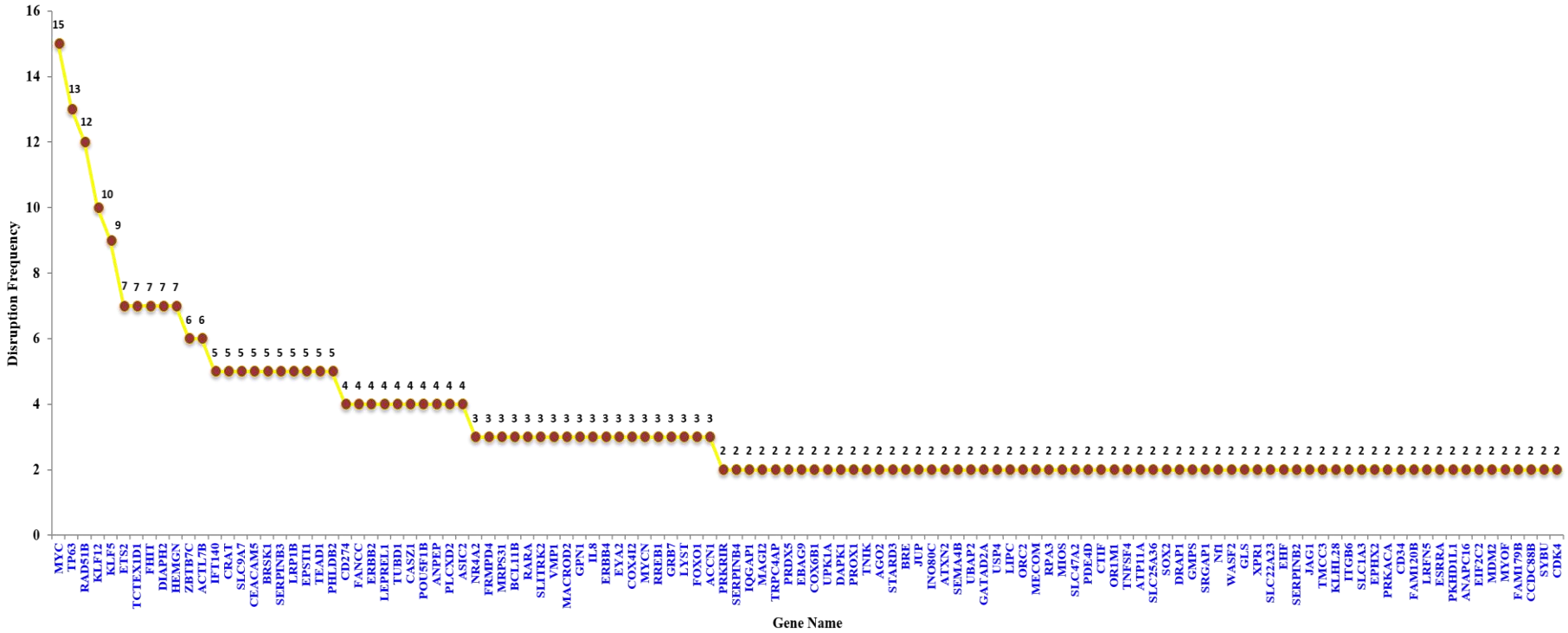


Figure 12. Depicting frequency of genes disrupted due to viral integration sites

HPV methylations

The quantitative status of HPV methylation from different diseased conditions was also cataloged in the study. These could be utilized and explored to discriminate cancer progression. Numerous studies provide distinct representations of methylation status, therefore to streamline notions, we have categorized methylation levels into three groups low or hypo-methylation (HypoM), high or hyper-methylation (HyperM), and significantly hyper-methylation (HyperM$^{\#\#}$). In total, 719 methylation entries were compiled with related clinical information. This mainly includes the HPV gene region, CpG methylation sites, methylation status, detection method, sample size, sample type (**Figure 13**). These entries belong to the 5 HPV genotypes viz. HPV16 (495),18 (113), 45 (66), 31 (34), and 33 (11) (**Figure 14**). Various studies have shown the relation between CpG methylation and carcinogenesis. Further, the integration event also generally correlates with the enhanced DNA methylation. We have provided the distinct methylation pattern corresponding to histology, specimens, and detection approach used. Maximum sites of late genes (L1 and L2) show a conclusive and significantly hyper-methylation profile. Correspondingly, long control region (LCR) exhibit inconsistent pattern. This can offer a comprehensive basis to compare distinctive methylation profiles from distinct cancer conditions that may enable advancement in screening tests.

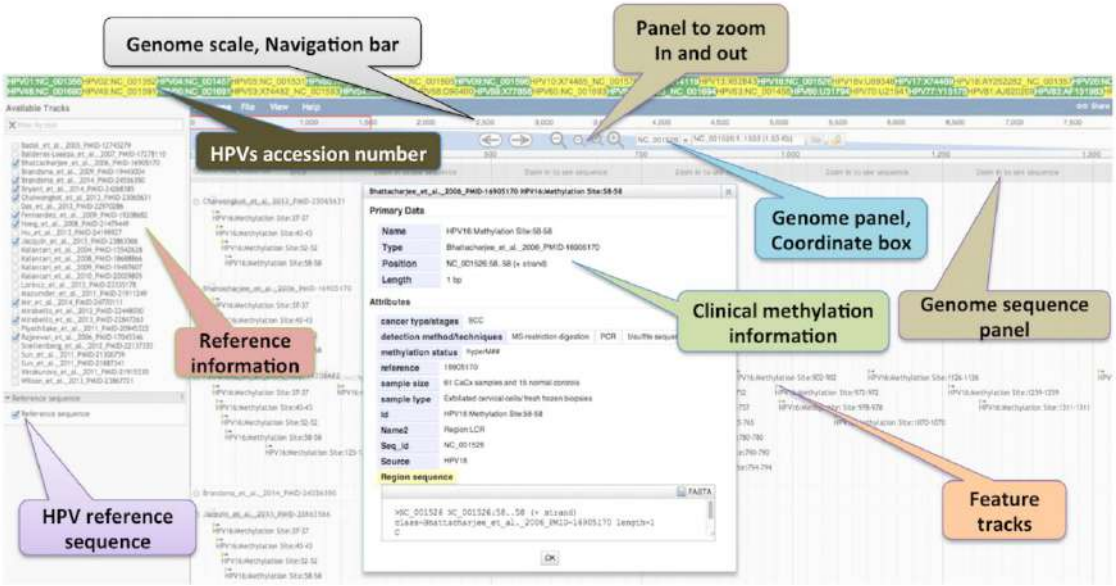


Figure 13. Screenshot showing highly interactive and user intensive methylation browser with associated histological information

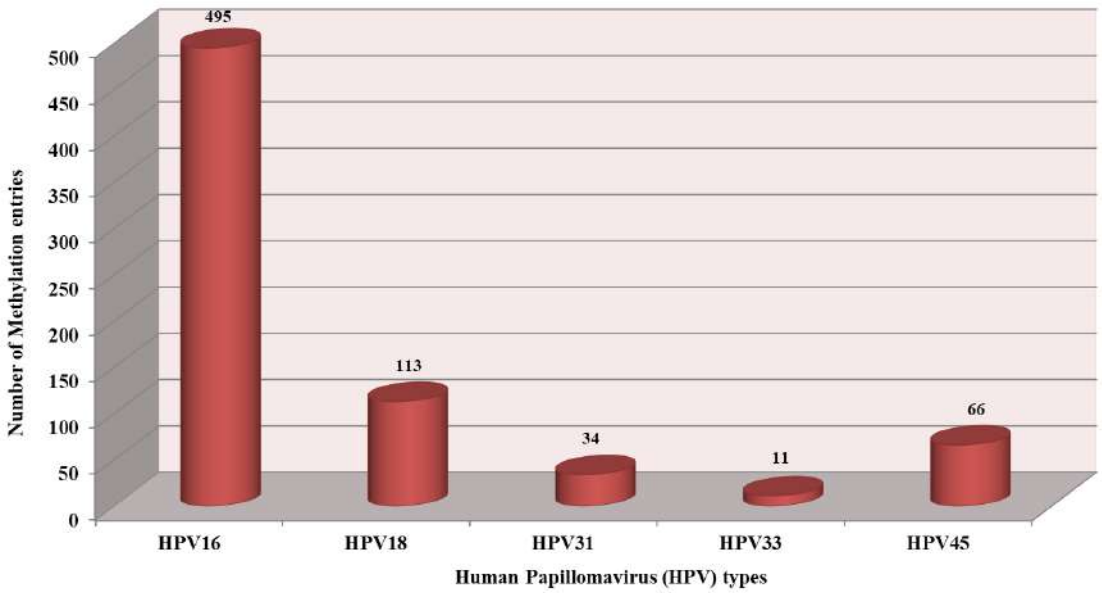


Figure 14. Graph showing distribution of methylation sites among distinct HPV types

Host miRNAs aberrant regulation

The abnormally (up or down) regulated cellular miRNAs are evident and known to be involved in the oncogenesis of diverse HPV-related cancers (Wang et al., 2014; Zhang et al., 2006). Thus, we have established a unified platform to facilitate an insightful analysis of aberrantly expressed miRNAs associated with HPV tumorigenesis. We have compiled and curated expression patterns of 341 miRNAs pertaining to different carcinomas, i.e., Cervical, head and neck, vulvar, and penile. This encompasses, 142 miRNAs covering both upregulated (80) and downregulated (62) from CaCx, 176 miRNAs comprise of 85 upregulated and 91 downregulated from HNSCC, 22 miRNAs (9 upregulated and 13 downregulated) belonging to vulvar carcinoma, and 1 miRNA (downregulated) from penile carcinoma. Relevant information like miRNAs, miRbase id, chromosome, genomic coordinates, and target genes was displayed (**Figure 15**). These miRNAs influence and target various host genes, which are mainly linked to cell proliferation, apoptosis, host defense system, senescence, metastasis, immune recognition, etc.

Besides, we have also explored the chromosomal distribution of these abruptly expressed miRNAs in distinct cancers namely, CaCx, HNSCC, and vulvar carcinoma. The chromosome wise distribution of these up and down-regulated miRNAs among these cancers is depicted in **Figure 16 and 17**. In CaCx, upregulated miRNAs are mostly found on chromosome X, 19, 1, 13, 7, 17, 5, etc. and downregulated are mainly distributed over chromosome 17, 19, 14, X, 1, and 9. Likewise, from HNSCC, chromosome X, 1, 9, 17, 13, and 7 harbor over-expressed miRNAs and chromosomes 9, 19, 14, 17, 1, X, 21, 3, and 11 mainly have under-expressed miRNAs. Correspondingly, from vulvar cancer mainly chromosome 17, 13, and 1 contain upregulated and chromosome 17 have downregulated miRNA.

Furthermore, cross regulated and intra-relationship of these miRNAs is also explored and analyzed. We have recognized the set of regularly over-expressed (**Figure 18a**) and under-expressed miRNAs (**Figure 18b**) in individual cancers. Additionally, it is important and interesting to notice that some miRNAs exhibit both (up and down) regulations in CaCx and HNSCC. In total, 23 miRNAs (**Figure 18c**) from HNSCC and 18 miRNAs (**Figure 18d**) from CaCx were identified that are reported to have both over as well as under expression. This could be a field of further research and exploration to get insights into the specific role of these in diverse conditions.

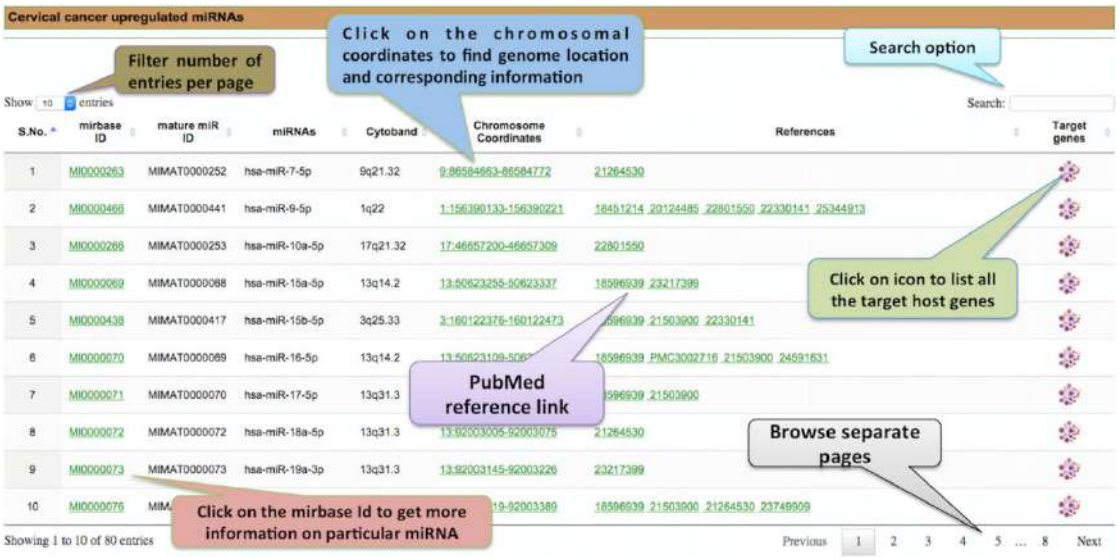
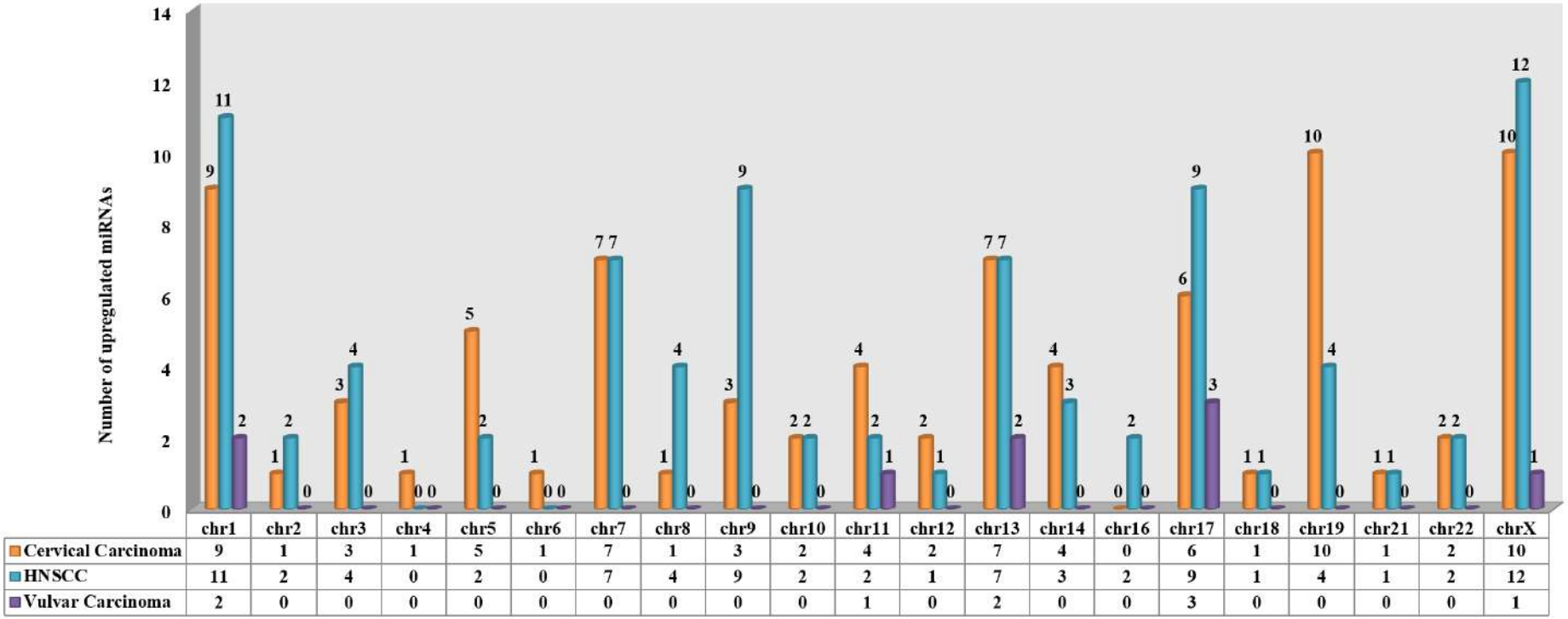
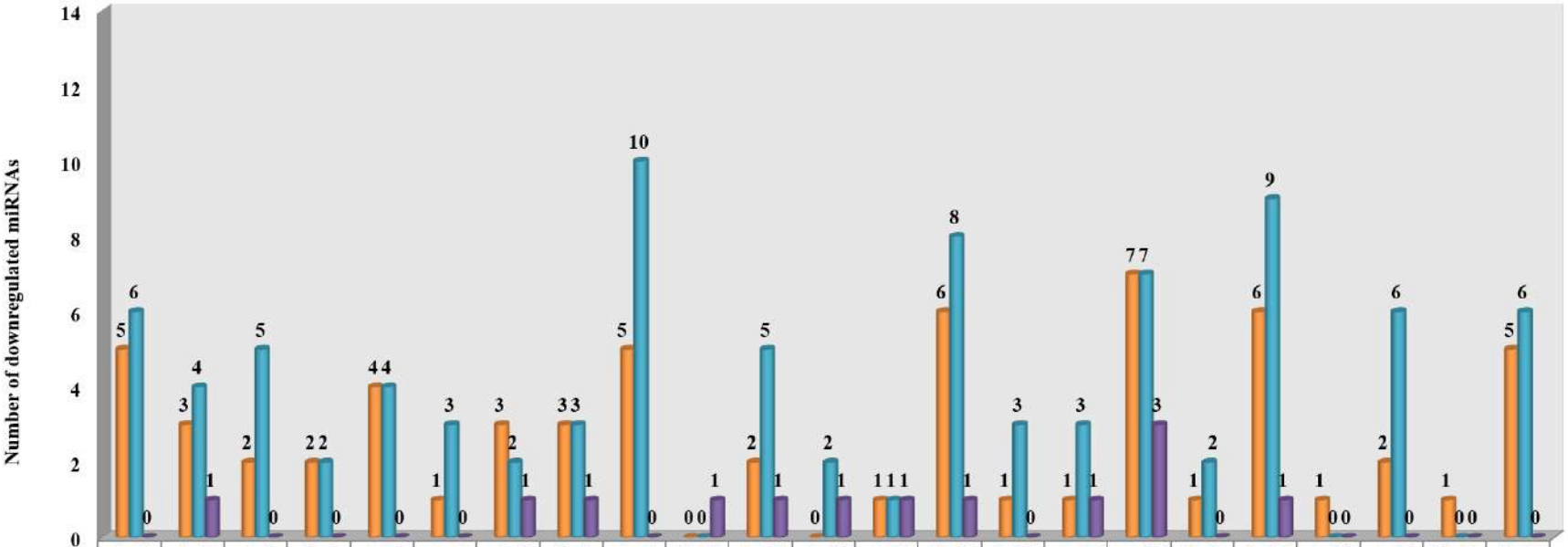


Figure 15. Screenshot illustrating HPV mediated upregulated miRNAs expression profile and analysis with interconnected external links



	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr16	chr17	chr18	chr19	chr21	chr22	chrX
Cervical Carcinoma	9	1	3	1	5	1	7	1	3	2	4	2	7	4	0	6	1	10	1	2	10
HNSCC	11	2	4	0	2	0	7	4	9	2	2	1	7	3	2	9	1	4	1	2	12
Vulvar Carcinoma	2	0	0	0	0	0	0	0	0	0	1	0	2	0	0	3	0	0	0	0	1

Figure 16. Plot showing chromosomal distribution of upregulated miRNAs in HPV associated carcinomas



	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX
Cervical Carcinoma	5	3	2	2	4	1	3	3	5	0	2	0	1	6	1	1	7	1	6	1	2	1	5
HNSCC	6	4	5	2	4	3	2	3	10	0	5	2	1	8	3	3	7	2	9	0	6	0	6
Vulvar Carcinoma	0	1	0	0	0	0	1	1	0	1	1	1	1	1	0	1	3	0	1	0	0	0	0

Figure 17. Plot showing chromosomal distribution of downregulated miRNAs in HPV associated carcinomas

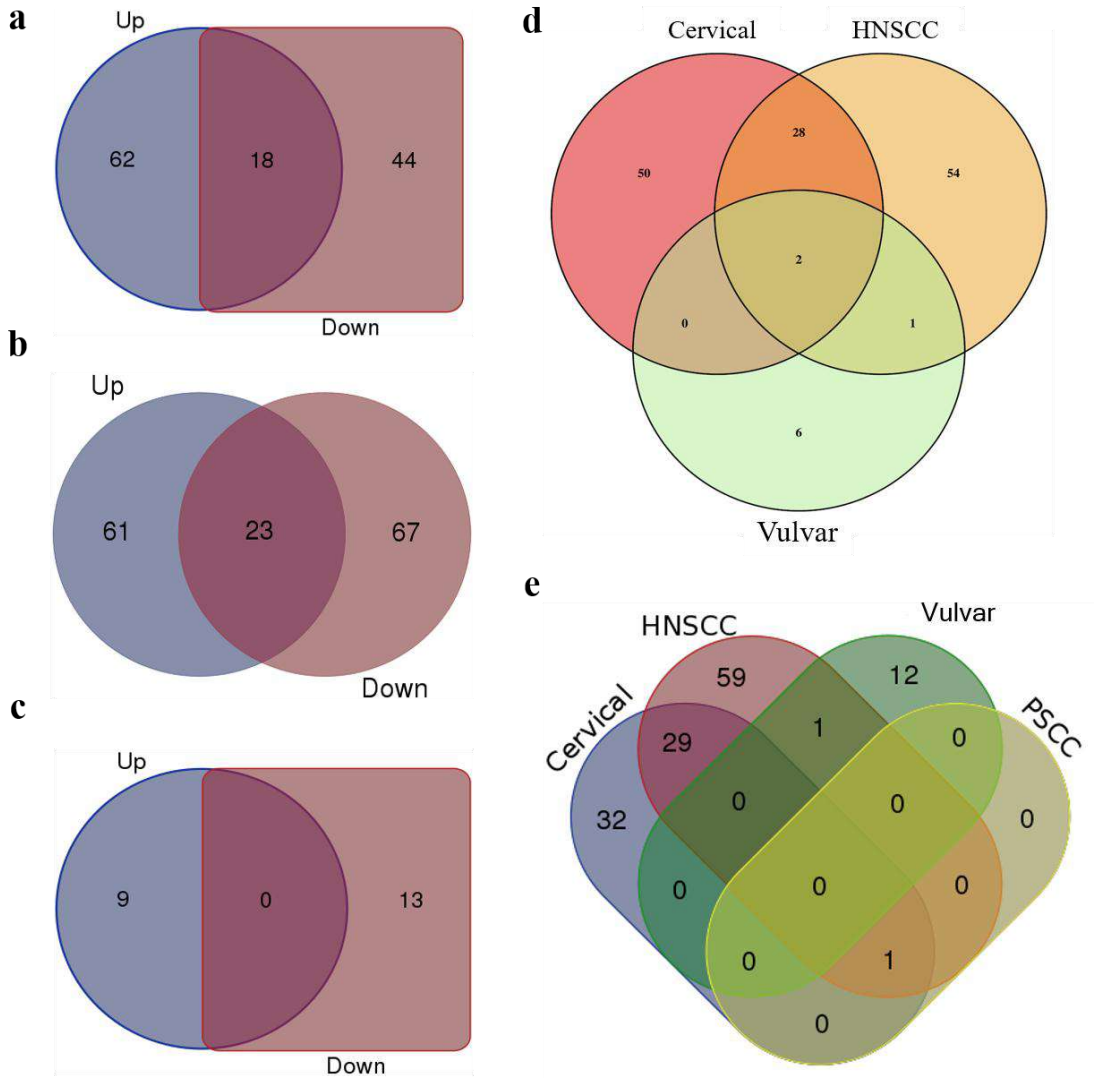


Figure 18. Figure illustrating commonly regulated miRNAs in diverse carcinomas. (a) Up and down regulated miRNAs in CaCx, (b) Up and down regulated miRNAs in HNSCC, and (c) Up and down regulated miRNAs in vulvar carcinoma (d) Upregulated miRNAs, (e) Downregulated miRNAs

Existing resources and comparison

Some resources were also developed to facilitate HPV genomic, proteomic, and epitope knowledge. This mainly includes the Papillomavirus Episteme (PaVE) (Van Doorslaer et al., 2013) and human papillomavirus T cell Antigen Database (HPVdb) (Zhang et al., 2014). However, there is no resource available for HPV mediated disease biomarkers. Additionally, there are also some relevant databases that are developed. Like, Dr.VIS is a database of disease linked viral integration sites that also have HPV related entries (Yang et al., 2015; Zhao et al., 2012a). Though, we have comprehensively analyzed and provided 1257 integration events with detailed associated knowledge. Similarly, for methylation, 8 databases were listed. These all are only specific to host methylation data (**Table 5**). Whereas, we have covered the methylome profile of HPVs from diverse carcinomas with related clinical information. Likewise, we also listed 19 miRNAs associated resources (**Table 6**). None of these provide abnormally regulated miRNAs pertinent to HPV infection. We cataloged 341 peculiar miRNAs with meta-information. These miRNAs signatures are valuable and can be important in the development of miRNA-based therapeutics. Like, the approach of reinstating or hindering miRNA functions can be employed in drug development. For example, a liposome-based miR-34 mimic (MRX34) is in clinical trial against hepatocellular carcinoma. Similarly, Miravirsen is an inhibitor of miR-122 biogenesis for the treatment of hepatitis C based on the locked nucleic acid (LNA) technology (Ling et al., 2013; Qureshi et al., 2014a).

Table 5. List of available major resources specific for methylation data

Database Name	References	Year
MethDB	(Amoreira et al., 2003)	2003
MethyCancer	(He et al., 2008)	2008
MethCancerDB	(Lauss et al., 2008)	2008
PubMeth	(Ongenaert et al., 2008)	2008
DiseaseMeth	(Lv et al., 2012)	2012
MethylomeDB	(Xin et al., 2012)	2012
PCMdb	(Nagpal et al., 2014)	2014
NGSmethDB	(Geisen et al., 2014)	2014

Table 6. List of available major resources related to microRNA data

Database Name	References	Year
miRGen	(Megraw et al., 2007)	2007
ViTa	(Hsu et al., 2007)	2007
miRNAMap	(Hsu et al., 2008)	2008
MicroRNA.org	(Betel et al., 2008)	2008
Vir-Mir db	(Li et al., 2008)	2008
miRecords	(Xiao et al., 2009)	2009
miROrtho	(Gerlach et al., 2009)	2009
miR2Disease	(Jiang et al., 2009)	2009
PMRD	(Zhang et al., 2010)	2010
TransmiR	(Wang et al., 2010)	2010
PmiRKB	(Meng et al., 2011)	2011
mESAdb	(Kaya et al., 2011)	2011
TarBase	(Vergoulis et al., 2012)	2012
miREX	(Bielewicz et al., 2012)	2012
miRGator	(Cho et al., 2013)	2013
miRBase	(Kozomara and Griffiths-Jones, 2014)	2014
miRTarBase	(Hsu et al., 2014)	2014
miRNEST	(Szczesniak and Makalowska, 2014)	2014
HMDD	(Li et al., 2014)	2014
VIRmiRNA	(Qureshi et al., 2014a)	2014

Conclusion

HPV infection and related cancers are associated with the sequence of events and risk factors such as HPV integration, methylation profiles, regulation of cellular miRNAs, etc. These changes can be utilized as potential biomarkers. Collectively, these events could be exploited for the improvement of prevention, screening, and therapeutic strategies (Brandsma et al., 2014; Clarke et al., 2012; de Freitas et al., 2014; Ling et al., 2013; Patel et al., 2012; Wang et al., 2014). Here, the aim of HPVbase (Kumar Gupta and Kumar, 2015) is to provide a comprehensive platform for the same. It is the first knowledgebase to deliver manually curated and an interactive resource of clinically valuable viral and cellular biomarkers. It comprises 1257 integration events from distinct HPV types mainly 16 (954), 18 (216), 33 (33), and 45 (33) related to different histological circumstances. Correspondingly, it also contains 719 quantitative HPV DNA methylation entries pertaining to 5 HPV genotypes namely HPV 16 (495), HPV 18 (113), HPV45 (66), HPV 31 (34) and HPV 33 (11). Furthermore, the aberrant expression of 341 miRNAs from diverse carcinoma along with their target genes were curated and compiled that can be useful for miRNA-based therapeutics. We anticipate that HPVbase would assist the scientific community engaged in HPV research.

Systematic meta-analysis of human genes disrupted due to HPVs associated events

Chapter 3. Systematic meta-analysis of human genes disrupted due to HPVs associated events

Introduction

Human papillomaviruses (HPVs) are the double-stranded DNA (dsDNA) onco-viruses from the *Papillomaviridae* family. Based on the malignant risks, these are further divided into two subgroups, i.e., high-risk HPVs that are highly carcinogenic in nature and low-risk HPVs (LR-HPVs) mainly associated with benign warts and lesions (de Martel et al., 2020; Gupta and Kumar, 2020). Persistency of HR-HPVs infection is highly crucial in the cancer progression towards precancer and invasion (Gupta and Kumar, 2020; Munoz et al., 2003). HPVs are reported to cause distinct cancers such as cervical, oropharyngeal, penile, vulvar, vaginal and anal carcinomas, etc. (Bray et al., 2018; de Martel et al., 2020). Two HPV oncogenes, i.e., E6 and E7 are mainly responsible for HPV oncogenesis (Moody and Laimins, 2010). HR-HPV types usually 16 and 18are known to play a vital role in the progression and etiology of cervical cancer, and head and neck squamous cell carcinomas (HNSCCs) (de Martel et al., 2020; Moody and Laimins, 2010).

Cervical cancer is the fourth most common cancer in women (~570000 cases) and the leading cause of mortality in women globally reported by Global cancer observatory (GLOBOCAN-2018) from the International Agency for Research on Cancer (IARC) (Arbyn et al., 2020; Bray et al., 2018; de Martel et al., 2020). HPVs are responsible for almost all cervical cancer cases (de Martel et al., 2020; Olusola et al., 2019). However, screening for the presence of HPVs and vaccination programs reduced the occurrences of cervical and other HPV-mediated cancers (de Martel et al., 2020; Olusola et al., 2019). Cervical carcinogenesis mainly proceeds with the HPVs infection followed by persistency that progress towards precancer, and high-grade/invasive carcinoma (Schiffman and Wentzensen, 2013). Based on severity, cervical cancer is also characterized in cervical intraepithelial neoplasia (CIN) I, II and III (Schiffman and Wentzensen, 2013). To postulate the role of HPVs in cervical cancer, Dr. Harald zur Hausen awarded the Nobel Prize in medicine in 2008 (zur Hausen, 2002, 2009**)**. Likewise, HPVs are also attributable to the HNSCC mainly oropharyngeal cancers (~42000 cases) (Bray et al., 2018; de Martel et al., 2020) and become an independent risk factor in HNSCC (D'Souza and Dempsey, 2011).

Along with the HPV infection and multi-stage progression of HPV-attributable cancers, diverse events and consequences are interconnected. This mainly includes HPV integration events, deregulation of miRNAs, epigenetic modifications and genomic alterations. (Kumar Gupta and Kumar, 2015; Schiffman and Wentzensen, 2013; Tuna and Amos, 2017). One of the key events is HPV integration that contributes towards carcinogenesis (McBride and Warburton, 2017). Integration event leads to the disruption and alterations of host genes and enhances genomic instability (Hu et al., 2015; Kumar Gupta and Kumar, 2015; Schiffman and Wentzensen, 2013).

Various studies reported the distribution and role of HPVs in different carcinomas (de Martel et al., 2020; Ojesina et al., 2014; Parfenov et al., 2014; Stransky et al., 2011; Tuna and Amos, 2017). Simultaneously, high-throughput data from the various cohorts of different carcinomas were analyzed for the distinct genomic, transcriptomic and epigenomic events (Ojesina et al., 2014; Rusan et al., 2015; Stransky et al., 2011; Tuna and Amos, 2017). Like, the landscape of genomic alterations was analyzed in cervical carcinoma (Ojesina et al., 2014). The Cancer Genome Atlas (TCGA) Research Network provides integrated genomic and molecular characterization of 228 cervical cancer samples (2017). Stransky et. al. reveals the mutational spectrum among genes in HNSCC utilizing large-scale sequencing (Stransky et al., 2011).

Furthermore, several studies also describe data analysis mainly through differential expressed genes (DEGs) identification to elucidate potential targets in different cancers including cervical and HNSCC utilizing different bioinformatics approaches (Costa et al., 2018; Kori and Yalcin Arga, 2018; López-Cortés et al., 2020; Zhang et al., 2020). This assimilates distinct strategies like differentially expressed genes (RNAseq and microarray), protein-protein interactome (PPI), co-expression networks, GO and pathway enrichment, protein expression, genomic (mutational profile, copy number variations (CNVs), mRNA regulation) and epigenomic alterations (Costa et al., 2018; Fang and Zhang, 2017; Fang et al., 2017; López-Cortés et al., 2020; Zhang et al., 2017; Zhang et al., 2020).

In the study, considering the heterogenicity of HPV-mediated carcinomas, an integrative approach merging multi-omics analysis along with the network biology is employed to elucidate HPV oncogenesis with a focus on clinical data from cervical squamous cell carcinoma (CSCC) and head and neck squamous cell carcinoma

(HNSCC). Known HPV infection-associated candidate genes were analyzed (Carvalho-Silva et al., 2019). Importantly, genes disrupted due to integration events in HPV pathogenesis (Kumar Gupta and Kumar, 2015) are also included in the meta-analysis. Our findings related to potential key and core therapeutic targets and relevant process categories, hallmark molecular functions, enriched pathways, genomic alterations, and potential drugs could aid towards the acceleration of clinical biomarker discovery and therapeutic development for HPV-linked carcinomas.

Materials and Method

Selection of candidate target genes

We have retrieved all the candidates from the two sources. First, we have extracted the genes from the "Open Targets Platform" which provide disease-specific potential targets (Carvalho-Silva et al., 2019). Overall, 1520 (G1 list) targets associated with the HPV infection were obtained. Second, a set of genes was acquired from our previous resource "HPVbase", which includes genes disrupted due to the HPV integration events (Kumar Gupta and Kumar, 2015). Overall, 463 (G2 list) HPV-integration associated genes were obtained. We have combined the G1 and G2 list to catalog the final working set. 96 duplicates were found between G1 and G2 and removed. In total, 1887 candidate genes were utilized for further downstream analysis.

PPI-Network based prioritization of potential target genes

All the 1887 candidate genes were searched implementing the Search Tool for the Retrieval of Interacting Genes (STRING) database (Szklarczyk et al., 2019) to identify protein-protein interactome with the high confidence score cut-off of 0.7. Further, the obtained interacting proteins were analyzed utilizing the Cytoscape program (Shannon et al., 2003). Furthermore, to identify the key target (Hub) genes from the OncoHPV-PPI network, cytoHubba (Chin et al., 2014) analysis utilizing the four different algorithms, i.e., Degree, Edge Percolated Component (EPC), Maximum Neighbourhood Component (MNC) and EcCentricity is performed.

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

Further, GO analysis and pathway enrichment analysis was performed on the prioritized target genes to explore the systematic functional and genomic annotations and pathway information. This includes identifying the critical biological processes,

cellular components, molecular functions and significant pathways in which these target genes were involved. For this, g:profiler (Raudvere et al., 2019) was employed for the systematic analyses, which provide multiple data sources for the comprehensive illustrations. Stringent cut-off criteria of value <0.001(G: SCS) was utilized for both GO and KEGG (Kanehisa et al., 2012) pathways analysis.

Gene set enrichment analysis (GSEA)

Gene set enrichment analysis of the target genes was conducted to detect different gene families and define top functional categories through hallmarks gene sets from Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) with significant cut-off value, i.e., FDR <0.05.

Analysis of the mutational profile, and copy number variations (CNVs) of selected target genes among cervical and Head and Neck carcinoma

Further, genomic alterations (mutations and CNVs) among the key target genes were explored pertaining to The Cancer Genome Atlas (TCGA)-cervical and head and neck carcinoma. Data from Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/) hosted at National Cancer Institute (NCI) (Weinstein et al., 2013) and cBioPortal (https://www.cbioportal.org) (Cerami et al., 2012) was utilized for the comprehensive meta-analyses. Overall, TCGA-PanCancer Atlas (PCA) provide genomic data of 307 Cervical squamous cell carcinoma (CSCC) and Endocervical Adenocarcinoma (CESC) cases (2017) and 529 samples with Head and Neck Squamous Cell Carcinoma (HNSCC) (Briese et al., 2015; Pérez Sayáns et al., 2019). Integrative analysis and OncoGrid is produced utilizing GDC Data Analysis, Visualization, and Exploration (DAVE) Tools**.** Ranking and a list of genes with high genomic alterations were also established.

Potential drugs for different target genes

Target genes were also analyzed for the potential drugs targeting these proteins utilizing the Open target platform (Carvalho-Silva et al., 2019). The drug-target network is cataloged and presented. Simultaneously, protein-protein relationships were also explored utilizing the OmniPath DB combining the 115 databases (Türei et al., 2016).

Results

OncoHPV-PPI Network and prioritization of target genes

In total, 1520 HPV infections associated genes were obtained from "Open Targets Platform" and 463 HPV integration disrupted genes from "HPVBase". After removing 96 duplicates from the list, 1887 candidate genes were utilized for the PPI network analysis using the String database and Cytoscape. With the (high) confidence score cut-off of 0.7, the final PPI-network had 1879 nodes and 20735 edges (**Figure 19**). Different nodes represent proteins in the network and edges show associations among proteins. The average node degree is 22.1 and PPI enrichment p-value <1.0e-16 in the constructed network.

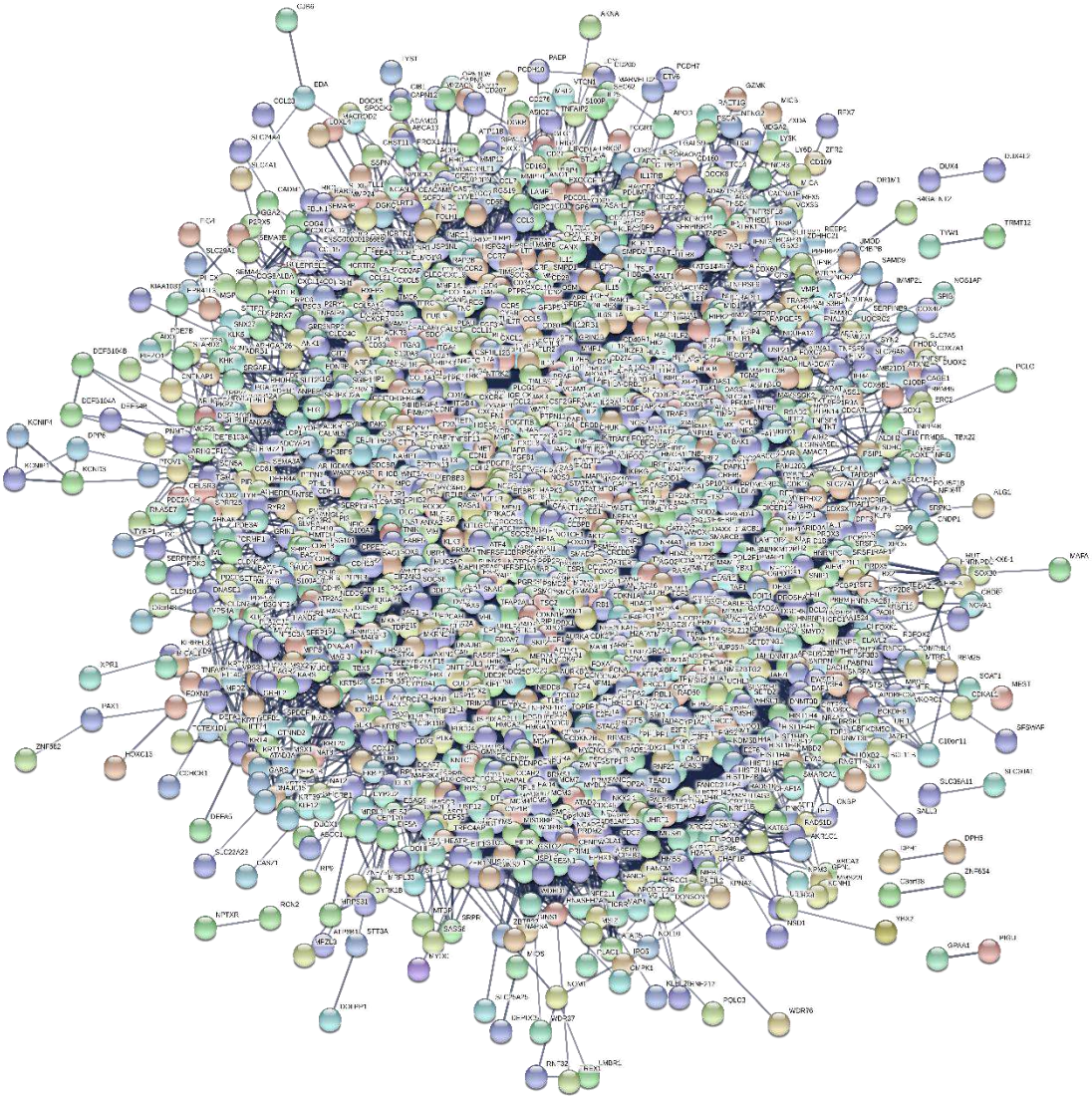


Figure 19. String based oncoHPV-PPI Network with 1879 nodes and 20735 edges with the high confidence score (>0.7)

All the connected nodes in oncoHPV-PPI Network were analyzed with CytoHubba application to explore the hub (target) genes. Top 100 genes from four different algorithms, i.e., Degree, Edge Percolated Component (EPC), Maximum Neighbourhood Component (MNC), and EcCentricity were retrieved (**Figure 20-23)**. Genes from each algorithm is integrated to deduce the oncoHPV-PPI core genes (**Figure 24 and Table 7**). Out of the 100 top genes from each algorithm, 44 genes were identified and named as the oncoHPV-PPI core genes. Further, taking degree as the significant criteria, the top 100 genes from the oncoHPV-PPI network (**Figure 20**) were utilized for the different downstream analyses.

Figure 20. Top 100 target genes in grid network with highest degrees and score from cytoHubba Degree algorithm

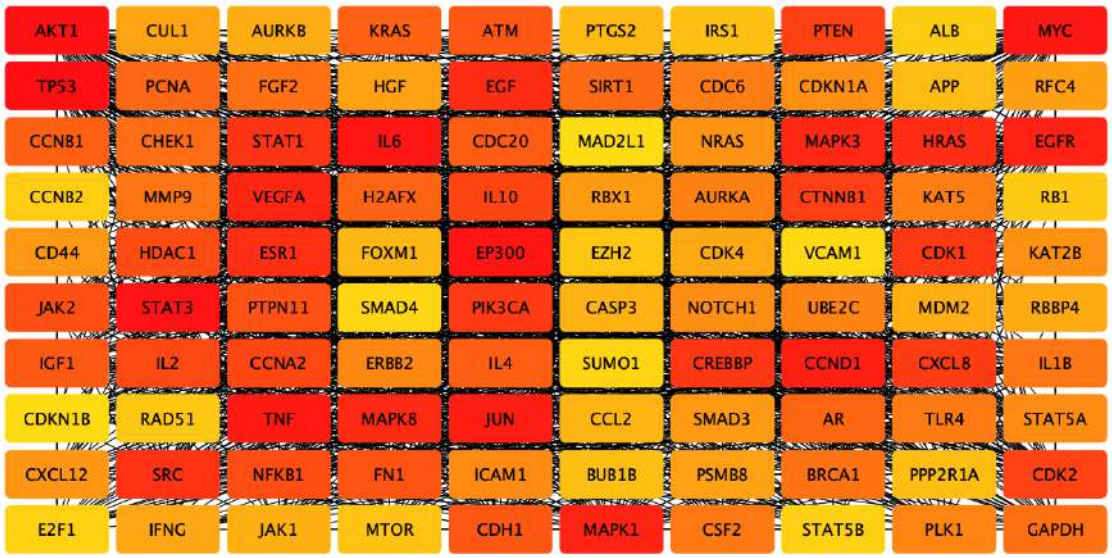


Figure 21. Grid network showing top 100 genes from Edge Percolated Component (EPC) method

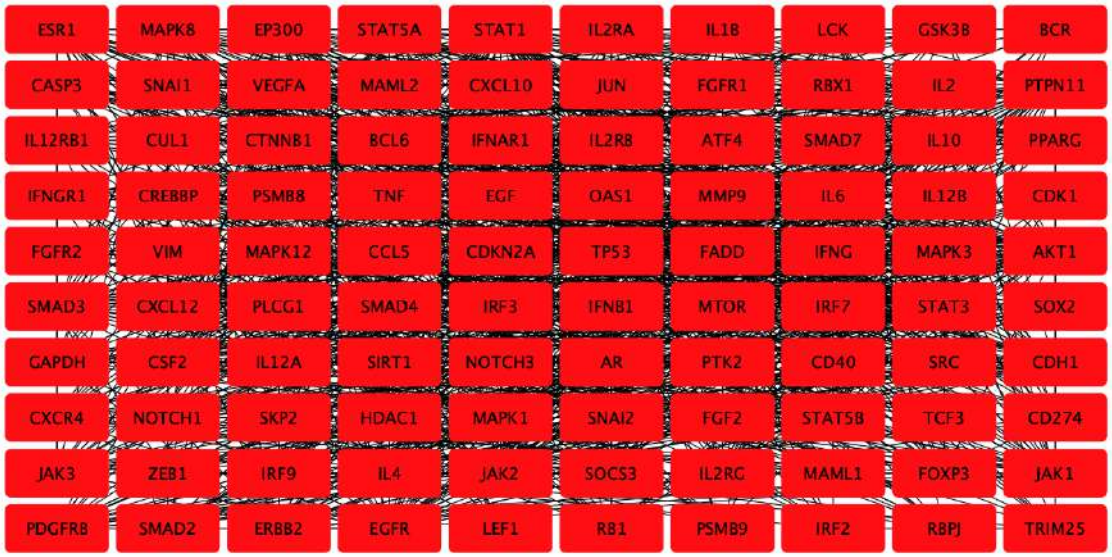


Figure 22. Grid network showing top 100 genes from EcCentricity algorithm

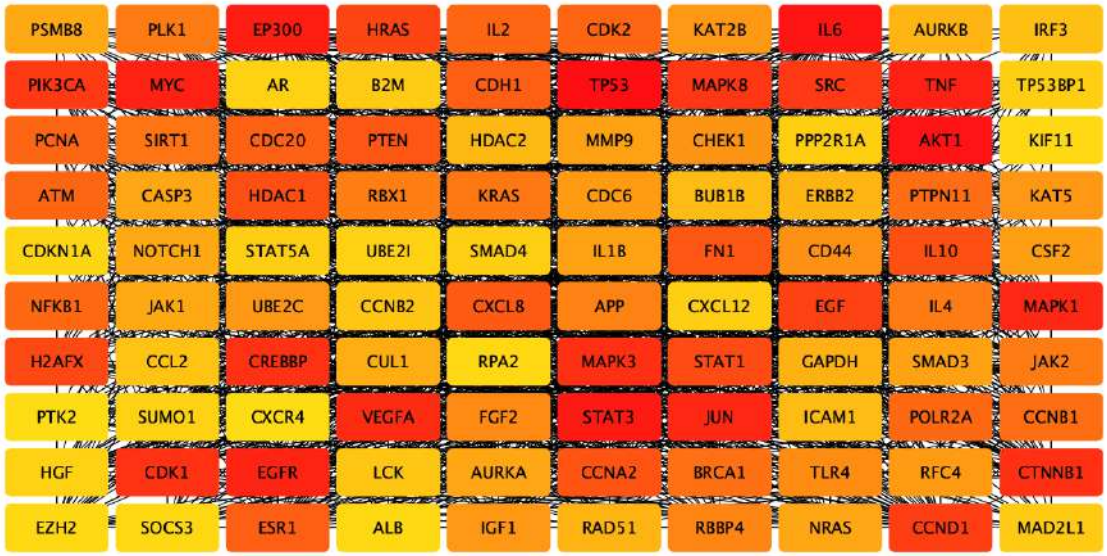


Figure 23. Grid network showing top 100 genes from Maximum Neighbourhood Component (MNC) method

Table 7. Table representing integration of significant targets from four different algorithms

Algorithms	*Genes names*
EcCentricity	VIM, SNAI2, IRF2, SMAD2, PPARG, IL12RB1, SNAI1, MAPK12, CD40, OAS1, IFNAR1, CXCL10, IRF7, ZEB1, CCL5, TRIM25, BCL6, CDKN2A, MAML2, RBPJ, SMAD7, IRF9, IL12A, IL12B, FADD, BCR, PDGFRB, CD274, PLCG1, IFNB1, NOTCH3, LEF1, SKP2, GSK3B, PSMB9, FOXP3, IL2RG, IL2RB, TCF3, MAML1, ATF4, FGFR1, SOX2, IFNGR1, FGFR2, IL2RA, JAK3
EPC	VCAM1, E2F1, CDK4, IRS1, CDKN1B, FOXM1, PTGS2
Degree and MNC	UBE2I, B2M, RPA2, KIF11, POLR2A, TP53BP1, HDAC2
MNC and EcCentricity	CXCR4
Degree and MNC and EcCentricity	IRF3, LCK, PTK2, SOCS3
Degree and MNC and EcCentricity and EPC	**EGFR, EGF, MAPK3, STAT3, AKT1, MMP9, AR, NOTCH1, SMAD4, ERBB2, CASP3, GAPDH, JUN, IL10, SMAD3, CXCL12, PSMB8, MAPK1, PTPN11, TP53, JAK1, CDK1, EP300, IL4, JAK2, STAT5A, CSF2, VEGFA, SRC, IL1B, CREBBP, HDAC1, IL6, CTNNB1, CUL1, CDH1, MAPK8, RBX1, STAT1, TNF, SIRT1, FGF2, ESR1, IL2**
Degree and MNC and EPC	**BRCA1, PTEN, SUMO1, FN1, HRAS, ALB, PIK3CA, CCL2, IGF1, APP, PCNA, CCNA2, TLR4, KRAS, RAD51, AURKA, CHEK1, MAD2L1, NFKB1, RFC4, CCND1, PLK1, H2AFX, CCNB1, MYC, KAT2B, CDC20, KAT5, HGF, AURKB, CCNB2, CDKN1A, CDK2, RBBP4, CXCL8, NRAS, CDC6, EZH2, PPP2R1A, ATM, CD44, ICAM1, BUB1B, UBE2C**

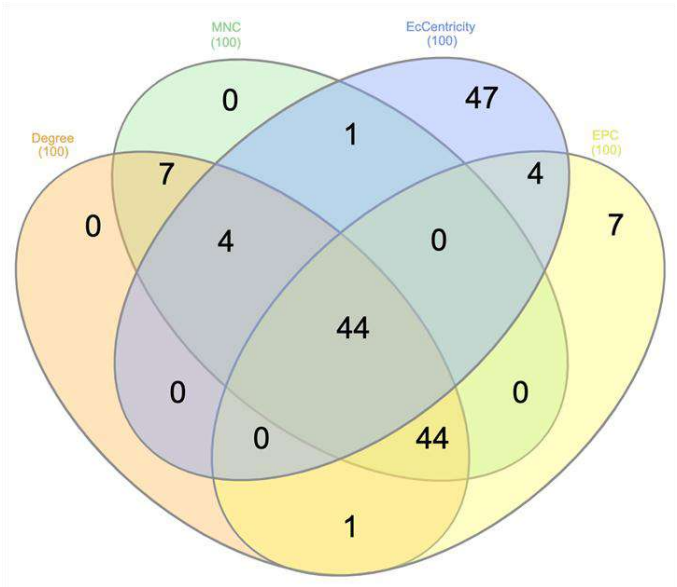


Figure 24. Venn diagram for integration of hub-genes (targets) from different algorithms, i.e., Degree, EPC, MNC and EcCentricity

Gene set enrichment analysis (GSEA)

Identified top 100 target genes were subjected to the different Gene Set Enrichment analysis (GSEA) investigated through Molecular Signatures Database (MSigDB) (Subramanian et al., 2005). Gene families and functional categories of these genes are cataloged. Overall, 70 genes were categorized in the 7 gene families, i.e., Oncogenes (21 genes), Transcription factors (TFs) (20), Protein kinases (19), Cytokines and growth factors (15), Translocated cancer genes (11), Tumor suppressors (8), and Cell differentiation (CD) markers (5). Target genes and corresponding gene families were shown using the Sankey plot (**Figure 25**).

Simultaneously, the top 20 enriched potential functional sets were identified (**Table 8**). The most significant three hallmarks functional sets among the target genes are G2M-Checkpoint (FDR q-value: 6.58E-25), E2F-Targets (3.60E-20), and Apoptosis (3.60E-20). G2M-Checkpoint includes 20 genes from targets. E2F-Targets includes 17 genes, Likewise, 16 target genes are marked under the Apoptosis hallmark. Some of the other important hallmarks also include mainly Signaling (TNFA_SIGNALING_VIA_NFKB, PI3K_AKT_MTOR_SIGNALING, IL6_JAK_STAT3_SIGNALING, WNT_BETA_CATENIN_SIGNALING and NOTCH_SIGNALING), MYC_TARGETS_V1 and P53_PATHWAY (**Table 8**).

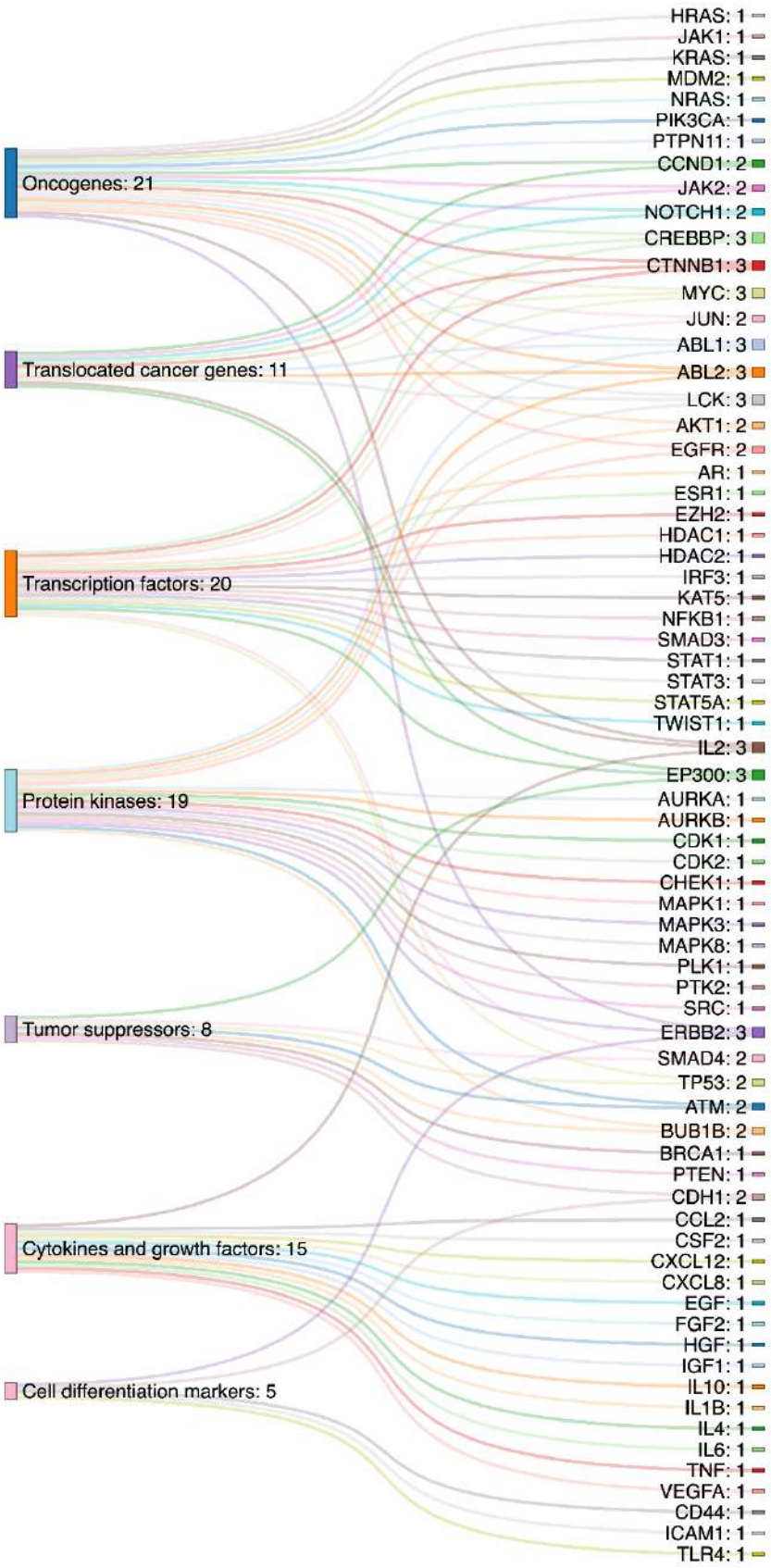


Figure 25. Different functional categories and gene families among target genes (Sankey plot)

Table 8. Hallmarks functions with genes from gene set enrichment analysis

Gene Set Name	Description	FDR q-value	Genes
HALLMARK_G2M_CHECKPOINT	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle.	6.58E-25	MYC, CDK1, CDC20, MAD2L1, RPA2, AURKA, CCNB2, EZH2, H2AX, PLK1, CHEK1, AURKB, CCND1, SMAD3, CUL1, CCNA2, CDC6, KIF11, EGF, UBE2C
HALLMARK_E2F_TARGETS	Genes encoding cell cycle related targets of E2F transcription factors.	3.60E-20	MYC, CDK1, CDC20, MAD2L1, RPA2, AURKA, CCNB2, EZH2, H2AX, PLK1, CHEK1, AURKB, BRCA1, CDKN1A, TP53, PCNA, BUB1B
HALLMARK_APOPTOSIS	Genes mediating programmed cell death (apoptosis) by activation of caspases.	3.60E-20	CCND1, BRCA1, CDKN1A, IL6, IL1B, TNF, CD44, JUN, CDK2, CASP3, CTNNB1, PTK2, APP, ERBB2, CREBBP, HGF
HALLMARK_ALLOGRAFT_REJECTION	Genes up-regulated during transplant rejection.	7.51E-19	BRCA1, IL6, IL1B, TNF, ICAM1, CCL2, LCK, EGFR, AKT1, IL4, STAT1, JAK2, B2M, IL10, MMP9, IL2
HALLMARK_TNFA_SIGNALING_VIA_NFKB	Genes regulated by NF-kB in response to TNF [GeneID=7124].	7.51E-19	MYC, CCND1, SMAD3, CDKN1A, IL6, IL1B, TNF, CD44, JUN, ICAM1, CCL2, SOCS3, CSF2, NFKB1, VEGFA, STAT5A
HALLMARK_PI3K_AKT_MTOR_SIGNALING	Genes up-regulated by activation of the PI3K/AKT/mTOR pathway.	3.96E-16	CDK1, CDKN1A, CDK2, LCK, EGFR, AKT1, IL4, PTPN11, HRAS, PTEN, MAPK1, MAPK8

Gene Set Name	Description	FDR q-value	Genes
HALLMARK_IL6_JAK_STAT3_SIGNALING	Genes up-regulated by IL6 [GeneID=3569] via STAT3 [GeneID=6774], e.g., during acute phase response.	1.44E-13	IL6, IL1B, TNF, CD44, JUN, STAT1, SOCS3, CSF2, PTPN11, STAT3
HALLMARK_INTERFERON_GAMMA_RESPONSE	Genes up-regulated in response to IFNG [GeneID=3458].	7.66E-13	CDKN1A, IL6, CASP3, ICAM1, CCL2, STAT1, JAK2, B2M, SOCS3, NFKB1, STAT3, PSMB8
HALLMARK_INFLAMMATORY_RESPONSE	Genes defining inflammatory response.	4.98E-10	MYC, CDKN1A, IL6, IL1B, ICAM1, CCL2, LCK, IL10, NFKB1, CXCL8
HALLMARK_WNT_BETA_CATENIN_SIGNALING	Genes up-regulated by activation of WNT signaling through accumulation of beta catenin CTNNB1 [GeneID=1499].	5.06E-09	MYC, CUL1, TP53, CTNNB1, HDAC2, NOTCH1
HALLMARK_MYC_TARGETS_V1	A subgroup of genes regulated by MYC - version 1 (v1).	9.37E-09	MYC, CDC20, MAD2L1, CUL1, CCNA2, PCNA, CDK2, HDAC2, RFC4
HALLMARK_APICAL_JUNCTION	Genes encoding components of apical junction complex.	1.40E-07	PTK2, ICAM1, EGFR, MMP9, HRAS, PTEN, CDH1, SRC
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	1.40E-07	IL6, CD44, JUN, VEGFA, CXCL8, CXCL12, FN1, FGF2
HALLMARK_ESTROGEN_RESPONSE_LATE	Genes defining late response to estrogen.	1.40E-07	CDC20, CCND1, CDC6, CD44, JAK2, CDH1, CXCL12, JAK1

Gene Set Name	Description	FDR q-value	Genes
HALLMARK_P53_PATHWAY	Genes involved in p53 pathways and networks.	1.40E-07	CDKN1A, TP53, PCNA, JUN, APP, HRAS, NOTCH1, MDM2
HALLMARK_DNA_REPAIR	Genes involved in DNA repair.	3.34E-07	RPA2, TP53, PCNA, RFC4, RBX1, POLR2A, RAD51
HALLMARK_COMPLEMENT	Genes encoding components of the complement system, which is part of the innate immune system.	2.21E-06	IL6, CASP3, LCK, JAK2, SRC, FN1, PIK3CA
HALLMARK_SPERMATOGENESIS	Genes up-regulated during production of male gametes (sperm), as in spermatogenesis.	3.27E-06	CDK1, AURKA, CCNB2, EZH2, RFC4, SIRT1
HALLMARK_NOTCH_SIGNALING	Genes up-regulated by activation of Notch signaling.	3.27E-06	CCND1, CUL1, NOTCH1, RBX1
HALLMARK_UV_RESPONSE_UP	Genes up-regulated in response to ultraviolet (UV) radiation.	7.50E-06	H2AX, IL6, CDK2, CASP3, ICAM1, RFC4

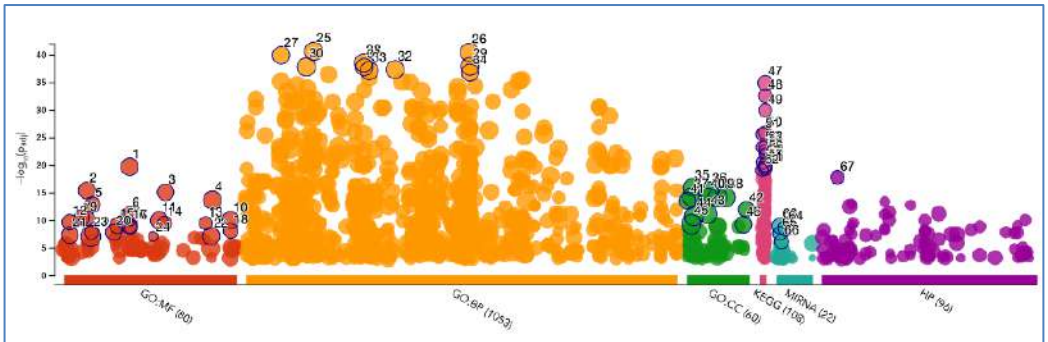
Gene ontology (GO) and KEGG pathways enrichment

Enrichment analysis for the oncoHPV-PPI top-100 genes (degree) is performed. g-profiler is employed to search significant molecular functions (MF), biological process (BP), cellular components (CC), KEGG biological pathways, regulatory motifs (miRNAs), and human disease phenotypes (HP) in Humans utilizing recommended "g:SCS" method with the stringent threshold of 0.001 (**Figure 26**). Overall, 80 molecular functions were determined, and the most significant GO: molecular function is enzyme binding (GO:0019899) with adjusted p-value (2.17E-20) and 53 interactions (EGFR, EGF, MAPK3, BRCA1, STAT3, PTEN, AKT1, AR, NOTCH1, ERBB2, LCK, CASP3, SUMO1, UBE2I, FN1, APP, JUN, PCNA, SMAD3, CCNA2, MAPK1, PTPN11, TP53, JAK1, MDM2, RAD51, AURKA, RFC4, CCND1, PLK1, CCNB1, JAK2, SRC, HDAC1, KAT2B, RPA2, CDC20, AURKB, CDKN1A, KIF11, PTK2, RBBP4, CTNNB1, CUL1, MAPK8, RBX1, CDC6, STAT1, TNF, SIRT1, ESR1, HDAC2, UBE2C). Significant molecular functions were depicted in **Figure 26**.

In total, 1053 biological process were deduced and the most significant GO: biological process is positive regulation of macromolecule metabolic process (GO:0010604) having 2.44E-41 adjusted p-value and 85 interactions (EGFR, EGF, MAPK3, BRCA1, STAT3, PTEN, AKT1, IRF3, MMP9, AR, NOTCH1, SMAD4, ERBB2, LCK, CASP3, SUMO1, UBE2I, FN1, GAPDH, HRAS, PIK3CA, CCL2, IGF1, APP, JUN, IL10, PCNA, SMAD3, CCNA2, MAPK1, PTPN11, TLR4, TP53, KRAS, MDM2, RAD51, CDK1, EP300, AURKA, CHEK1, NFKB1, RFC4, CCND1, PLK1, CCNB1, MYC, IL4, B2M, JAK2, STAT5A, CSF2, VEGFA, SRC, IL1B, CREBBP, HDAC1, KAT2B, CDC20, KAT5, HGF, AURKB, CDKN1A, PTK2, IL6, CDK2, CXCL8, CTNNB1, CDH1, MAPK8, RBX1, CDC6, STAT1, TNF, EZH2, TP53BP1, ATM, SIRT1, FGF2, CD44, ICAM1, SOCS3, ESR1, HDAC2, IL2, UBE2C). Substantial biological processes were shown in **Figure 26**.

Further, 108 biological pathways were identified and the most significant KEGG biological pathway was Pathways in cancer (KEGG:05200) with adjusted p-value (1.46E-35) and 51 interactions (EGFR, EGF, MAPK3, STAT3, PTEN, AKT1, MMP9, AR, NOTCH1, SMAD4, ERBB2, CASP3, FN1, HRAS, PIK3CA, IGF1, JUN, SMAD3, CCNA2, CXCL12, MAPK1, TP53, JAK1, KRAS, MDM2, RAD51, EP300, NFKB1, CCND1, MYC, IL4, JAK2, STAT5A, VEGFA, CREBBP, HDAC1, HGF, CDKN1A, PTK2, IL6, CDK2, CXCL8, CTNNB1, CUL1, CDH1, MAPK8, NRAS, RBX1, STAT1, FGF2, ESR1, HDAC2, IL2). Significantly enriched biological pathways are shown in **Figure 26**.

Likewise, 22 regulatory miRNAs were marked and the most significant is hsa-miR-155-5p (adjusted p-value: 1.12E-09) with 28 interactions (EGFR, STAT3, PTEN, AKT1, SMAD4, CASP3, PIK3CA, CCL2, JUN, SMAD3, KRAS, RAD51, AURKA, NFKB1, CCND1, PLK1, MYC, AURKB, IL6, CDK2, CXCL8, CTNNB1, STAT1, SIRT1, FGF2, ICAM1, SOCS3, IL2) (**Table 9**). Additionally, 96 human phenotypes were recognized and the most significant ontology was Somatic mutation (HP:0001428) with 1.68E-18 adjusted p-value (**Figure 26**) and 25 interactions (EGFR, BRCA1, PTEN, AKT1, AR, SMAD4, ERBB2, HRAS, PIK3CA, PTPN11, TP53, KRAS, RAD51, EP300, AURKA, CCND1, MYC, JAK2, SRC, CTNNB1, CDH1, NRAS, ATM, ESR1, BUB1B).



ID	Source	Term ID	Term Name	P_{adj} (query_1)
1	GO:MF	GO:0019899	enzyme binding	2.177×10^{-20}
2	GO:MF	GO:0005102	signaling receptor binding	4.341×10^{-16}
3	GO:MF	GO:0044877	protein-containing complex binding	8.687×10^{-16}
4	GO:MF	GO:0098772	molecular function regulator	2.313×10^{-14}
5	GO:MF	GO:0008134	transcription factor binding	1.631×10^{-13}
6	GO:MF	GO:0019900	kinase binding	8.022×10^{-12}
7	GO:MF	GO:0004672	protein kinase activity	3.404×10^{-11}
8	GO:MF	GO:0019901	protein kinase binding	4.301×10^{-11}
9	GO:MF	GO:0005126	cytokine receptor binding	4.905×10^{-11}
10	GO:MF	GO:0140096	catalytic activity, acting on a protein	1.121×10^{-10}
11	GO:MF	GO:0042802	identical protein binding	1.167×10^{-10}
12	GO:MF	GO:0003682	chromatin binding	2.356×10^{-10}
13	GO:MF	GO:0070851	growth factor receptor binding	3.597×10^{-10}
14	GO:MF	GO:0044389	ubiquitin-like protein ligase binding	3.849×10^{-10}
15	GO:MF	GO:0016773	phosphotransferase activity, alcohol group as acceptor	8.871×10^{-10}
16	GO:MF	GO:0019902	phosphatase binding	1.346×10^{-9}
17	GO:MF	GO:0019904	protein domain specific binding	1.397×10^{-9}
18	GO:MF	GO:0140297	DNA-binding transcription factor binding	5.648×10^{-9}
19	GO:MF	GO:0008022	protein C-terminus binding	1.701×10^{-8}
20	GO:MF	GO:0016301	kinase activity	1.711×10^{-8}
21	GO:MF	GO:0003690	double-stranded DNA binding	5.320×10^{-8}
22	GO:MF	GO:0097367	carbohydrate derivative binding	7.681×10^{-8}
23	GO:MF	GO:0005515	protein binding	9.022×10^{-8}
24	GO:MF	GO:0035173	histone kinase activity	8.888×10^{-8}
25	GO:BP	GO:0010604	positive regulation of macromolecule metabolic process	2.442×10^{-41}
26	GO:BP	GO:0051173	positive regulation of nitrogen compound metabolic process	3.989×10^{-41}
27	GO:BP	GO:0006468	protein phosphorylation	1.336×10^{-40}
28	GO:BP	GO:0031325	positive regulation of cellular metabolic process	3.093×10^{-39}
29	GO:BP	GO:0051246	regulation of protein metabolic process	1.345×10^{-38}
30	GO:BP	GO:0009893	positive regulation of metabolic process	1.672×10^{-38}
31	GO:BP	GO:0031399	regulation of protein modification process	1.770×10^{-36}
32	GO:BP	GO:0035556	intracellular signal transduction	4.905×10^{-36}
33	GO:BP	GO:0032268	regulation of cellular protein metabolic process	8.330×10^{-36}
34	GO:BP	GO:0051247	positive regulation of protein metabolic process	1.734×10^{-27}
35	GO:CC	GO:0005694	chromosome	1.163×10^{-16}
36	GO:CC	GO:0032991	protein-containing complex	6.004×10^{-16}
37	GO:CC	GO:0005654	nucleoplasm	6.229×10^{-15}
38	GO:CC	GO:0070013	intracellular organelle lumen	7.471×10^{-15}
39	GO:CC	GO:0043233	organelle lumen	7.471×10^{-15}
40	GO:CC	GO:0031974	membrane-enclosed lumen	7.471×10^{-15}
41	GO:CC	GO:0000228	nuclear chromosome	3.200×10^{-14}
42	GO:CC	GO:1990234	transferase complex	1.171×10^{-12}
43	GO:CC	GO:0031981	nuclear lumen	7.534×10^{-12}
44	GO:CC	GO:0005829	cytosol	2.072×10^{-11}
45	GO:CC	GO:0005634	nucleus	5.282×10^{-10}
46	GO:CC	GO:1902494	catalytic complex	6.230×10^{-10}
47	KEGG	KEGG:05200	Pathways in cancer	1.456×10^{-35}
48	KEGG	KEGG:05166	Human T-cell leukemia virus 1 infection	2.939×10^{-33}
49	KEGG	KEGG:05161	Hepatitis B	1.243×10^{-30}
50	KEGG	KEGG:04933	AGE-RAGE signaling pathway in diabetic complications	1.317×10^{-28}
51	KEGG	KEGG:04068	FoxO signaling pathway	2.626×10^{-26}
52	KEGG	KEGG:04110	Cell cycle	6.050×10^{-24}
53	KEGG	KEGG:05167	Kaposi sarcoma-associated herpesvirus infection	7.034×10^{-24}
54	KEGG	KEGG:05205	Proteoglycans in cancer	4.500×10^{-23}
55	KEGG	KEGG:05219	Bladder cancer	8.758×10^{-23}
56	KEGG	KEGG:05215	Prostate cancer	2.661×10^{-22}
57	KEGG	KEGG:04218	Cellular senescence	3.305×10^{-21}
58	KEGG	KEGG:05163	Human cytomegalovirus infection	1.022×10^{-20}
59	KEGG	KEGG:05203	Viral carcinogenesis	1.050×10^{-20}
60	KEGG	KEGG:05212	Pancreatic cancer	3.405×10^{-20}
61	KEGG	KEGG:05165	Human papillomavirus infection	3.525×10^{-20}
62	KEGG	KEGG:01521	EGFR tyrosine kinase inhibitor resistance	8.054×10^{-20}
63	MIRNA	MIRNA:hsa-miR-155-5p	hsa-miR-155-5p	1.116×10^{-9}
64	MIRNA	MIRNA:hsa-miR-34a-5p	hsa-miR-34a-5p	2.774×10^{-8}
65	MIRNA	MIRNA:hsa-miR-146a-5p	hsa-miR-146a-5p	2.856×10^{-8}
66	MIRNA	MIRNA:hsa-miR-203a-3p	hsa-miR-203a-3p	7.622×10^{-7}
67	HP	HP:0001428	Somatic mutation	1.680×10^{-16}

Figure 26. Enrichment map of GO: MF, GO: BP, GO: CC, KEGG pathways, miRNAs, and the human phenotype ontology. Most significant features in each category is marked and listed in the figure

Mutational profile and copy number variation profile among cervical and HNSCC

TCGA (https://www.cancer.gov/tcga) present the clinical data of 307 cases from the Cervix related to Cervical squamous cell carcinoma (CSCC) and Endocervical Adenocarcinoma (CESC project) and 529 cases from Head and Neck Squamous Cell Carcinoma (HNSCC). Genomic alterations, i.e., mutational and Copy Number Variation (CNV) profiles of target genes were analyzed for both the carcinomas.

Out of total 307 CESC cases, 293 samples show mutation (280 cases) or CNV data (284 cases) for the target genes. OncoGrid of genomic alterations in CESC provides a comprehensive view of 50 frequently mutated genes with CNVs (loss and gain) among samples (**Figure 27**). The top 20 most frequently mutated genes are presented in **Figure 28**. The most frequently mutated genes in the CESC among selected targets are PIK3CA (3q26.32), EP300 (22q13.2), PTEN (10q23.31), CREBBP (16p13.3), NOTCH1 (9q34.3), TP53 (17p13.1), KRAS (12p12.1), ERBB2 (17q12), TP53BP1 (15q15.3), SMAD4 (18q21.2), AR (Xq12), MAPK1 (22q11.22), POLR2A (17p13.1), EGFR (7p11.2), BRCA1 (17q21.31), ATM (11q22.3), FN1 (2q35), ESR1 (6q25.1, 6q25.2), CUL1 (7q36.1), and PPP2R1A (19q13.41) (**Figure 28**). In total, 860 mutations pertaining to target genes were marked. Most mutations are present on EP300, PTEN, NOTCH1 CREBBP, PIK3CA, and TP53 (**Figure 29**). The top 20 genes with the highest number of mutations are depicted in **Figure 29**. Likewise, genes with high CNVs (gain and loss) are shown in **Figure 30-31**.

Table 9. Gene ontology based 22 miRNA targets

MiRNAs	*Adjusted p_value*	*Intersection size*	*Intersections*
hsa-miR-155-5p	1.12E-09	28	EGFR, STAT3, PTEN, AKT1, SMAD4, CASP3, PIK3CA, CCL2, JUN, SMAD3, KRAS, RAD51, AURKA, NFKB1, CCND1, PLK1, MYC, AURKB, IL6, CDK2, CXCL8, CTNNB1, STAT1, SIRT1, FGF2, ICAM1, SOCS3, IL2
hsa-miR-34a-5p	2.77E-09	25	MAPK3, BRCA1, AKT1, AR, NOTCH1, SMAD4, ERBB2, CASP3, IL10, TP53, RAD51, NFKB1, CCND1, MYC, VEGFA, SRC, HDAC1, CDC20, KIF11, CTNNB1, STAT1, POLR2A, TNF, SIRT1, CD44
hsa-miR-146a-5p	2.66E-08	14	EGFR, BRCA1, NOTCH1, SMAD4, CCNA2, CXCL12, TLR4, NFKB1, CCND1, CDKN1A, IL6, CXCL8, STAT1, ICAM1
hsa-miR-203a-3p	7.62E-07	15	SMAD4, SUMO1, PIK3CA, JUN, VEGFA, SRC, IL6, CXCL8, CDH1, MAPK8, STAT1, TNF, ATM, FGF2, SOCS3
hsa-miR-92a-3p	1.65E-06	30	STAT3, PTEN, SMAD4, GAPDH, MDM2, RAD51, CDK1, EP300, AURKA, CHEK1, NFKB1, CCND1, CCNB1, MYC, HDAC1, KAT2B, RPA2, CDC20, AURKB, CDH1, MAPK8, NRAS, RBX1, CDC6, PPP2R1A, ATM, SIRT1, FGF2, ICAM1, HDAC2
hsa-miR-193b-3p	2.23E-06	23	BRCA1, PTEN, AKT1, PCNA, SMAD3, CCNA2, PTPN11, KRAS, RAD51, CDK1, EP300, CHEK1, RFC4, CCND1, CDC20, KIF11, CDH1, MAPK8, CDC6, EZH2, ESR1, BUB1B, UBE2C
hsa-miR-26a-5p	2.48E-05	16	BRCA1, PTEN, SMAD4, IGF1, MDM2, EP300, CHEK1, MAD2L1, MYC, HGF, IL6, NRAS, CDC6, EZH2, ATM, ESR1
hsa-miR-145-5p	2.75E-05	12	EGFR, SMAD4, SMAD3, MDM2, MYC, VEGFA, CDKN1A, NRAS, STAT1, CD44, ESR1, HDAC2

MiRNAs	*Adjusted p_value*	*Intersection size*	*Intersections*
hsa-miR-24-3p	5.78E-05	21	BRCA1, NOTCH1, SUMO1, CCL2, IGF1, PCNA, CCNA2, TP53, CDK1, AURKA, CHEK1, CCND1, CCNB1, MYC, IL4, IL1B, HDAC1, AURKB, RBBP4, TNF, UBE2C
hsa-miR-199a-5p	6.95E-05	10	SMAD4, ERBB2, SMAD3, KRAS, NFKB1, VEGFA, CDH1, EZH2, SIRT1, CD44
hsa-miR-223-3p	0.000136322	8	STAT3, TP53, MDM2, STAT5A, IL6, CDK2, STAT1, ATM
hsa-miR-429	0.000375637	9	PTEN, JUN, KRAS, EP300, MYC, IL4, VEGFA, RBBP4, EZH2
hsa-let-7a-5p	0.000432005	17	EGFR, STAT3, CASP3, SUMO1, HRAS, APP, KRAS, NFKB1, CCND1, MYC, AURKB, CCNB2, CDKN1A, IL6, CXCL8, NRAS, EZH2
hsa-miR-199a-3p	0.00047403	8	AKT1, IGF1, MAPK1, VEGFA, HGF, MAPK8, FGF2, CD44
hsa-miR-30a-5p	0.000612398	18	EGFR, NOTCH1, CASP3, JUN, PCNA, MAPK1, TP53, JAK1, HDAC1, RPA2, CDC20, KIF11, CTNNB1, CDH1, MAPK8, ATM, CD44, SOCS3
hsa-miR-886-3p	0.000658001	3	CXCL12, PLK1, CDC6
hsa-miR-22-3p	0.000713915	9	PTEN, AKT1, ERBB2, CCNA2, PLK1, CDKN1A, SIRT1, ESR1, BUB1B
hsa-miR-25-3p	0.000734903	15	PTEN, ERBB2, GAPDH, TP53, MDM2, RAD51, EP300, AURKA, CCNB1, MYC, KAT2B, CDH1, NRAS, EZH2, FGF2
hsa-miR-200c-3p	0.000813288	10	PTEN, NOTCH1, UBE2I, FN1, JUN, KRAS, EP300, VEGFA, CDK2, SIRT1
hsa-miR-125a-5p	0.000868411	11	EGFR, STAT3, AKT1, SMAD4, ERBB2, TP53, MYC, JAK2, VEGFA, CDKN1A, MAPK8
hsa-miR-138-5p	0.00095525	8	AKT1, CASP3, NFKB1, CCND1, PTK2, CDH1, EZH2, SIRT1
hsa-miR-451a	0.000975651	5	AKT1, MMP9, MAPK1, MYC, IL6

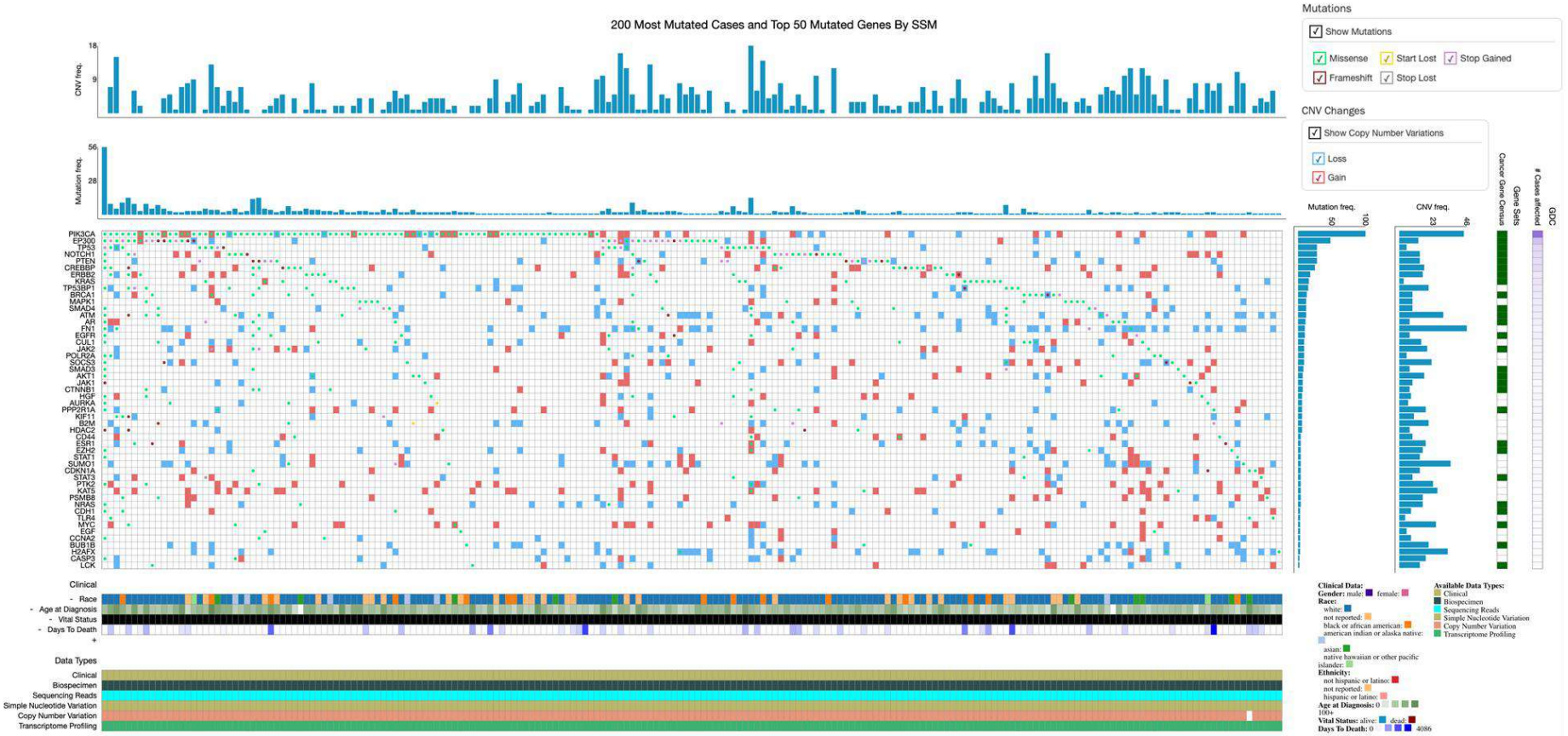


Figure 27. OncoGrid presenting top 50 genes with genomic alterations (mutations and CNVs) in CESC

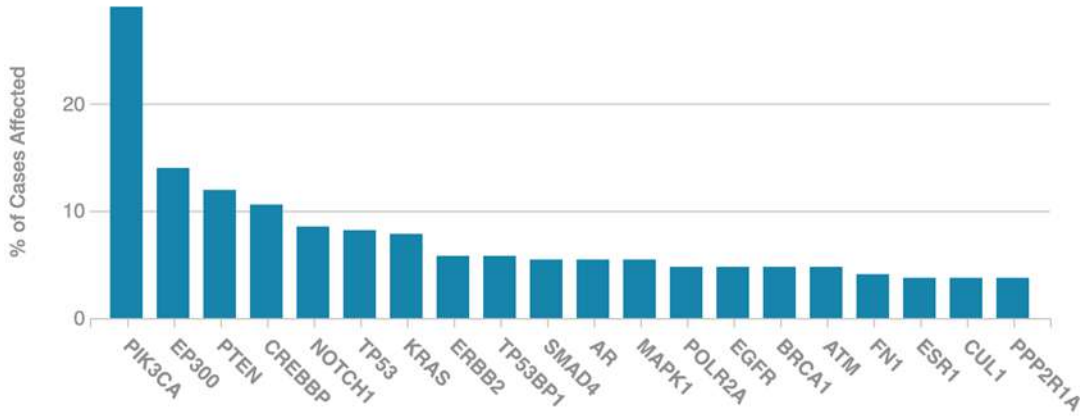


Figure 28. The most frequently mutated genes among targets in CESC samples

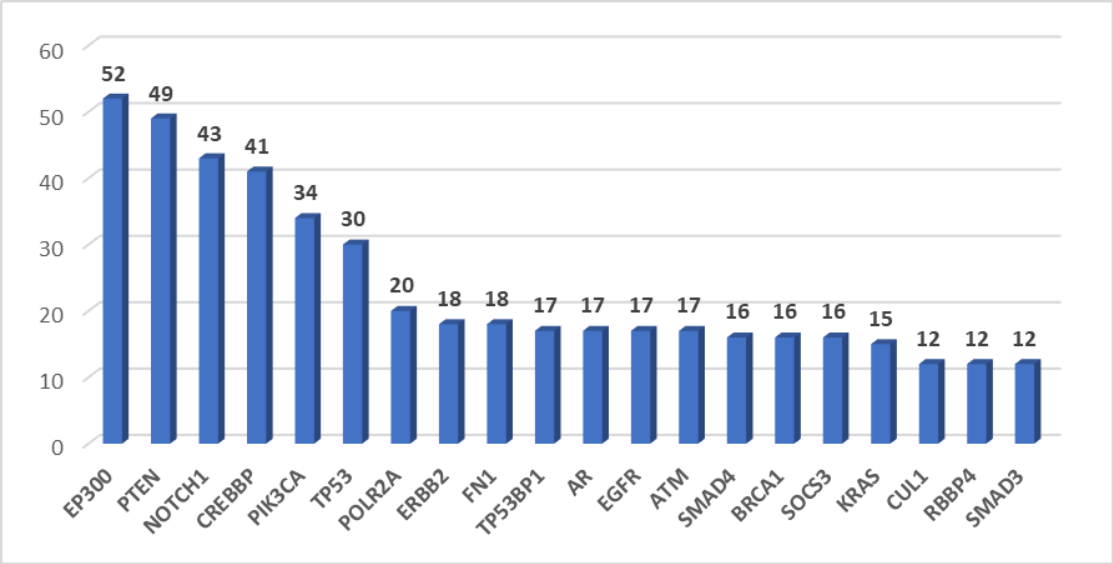


Figure 29. The most mutated target genes and number of mutations in CESC samples

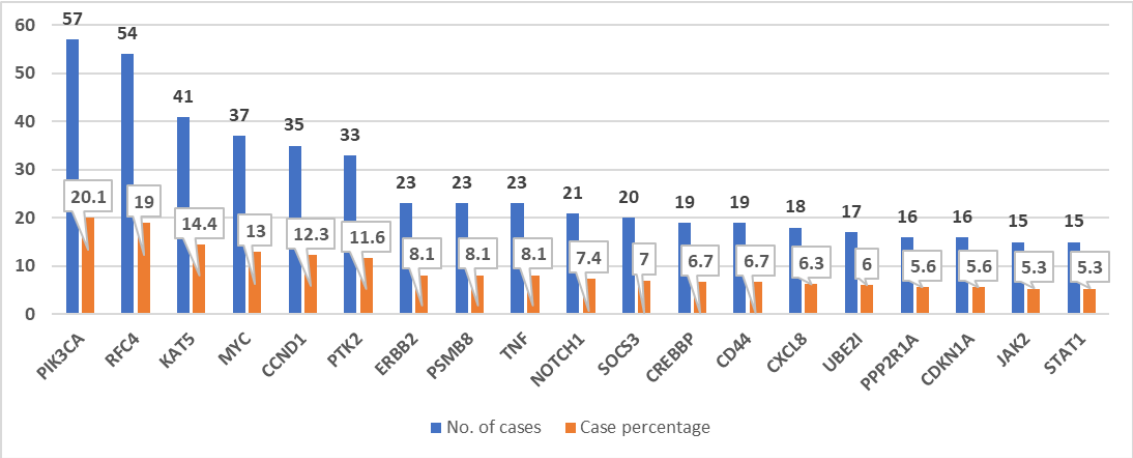


Figure 30. Target genes with significant copy number gain in CESC

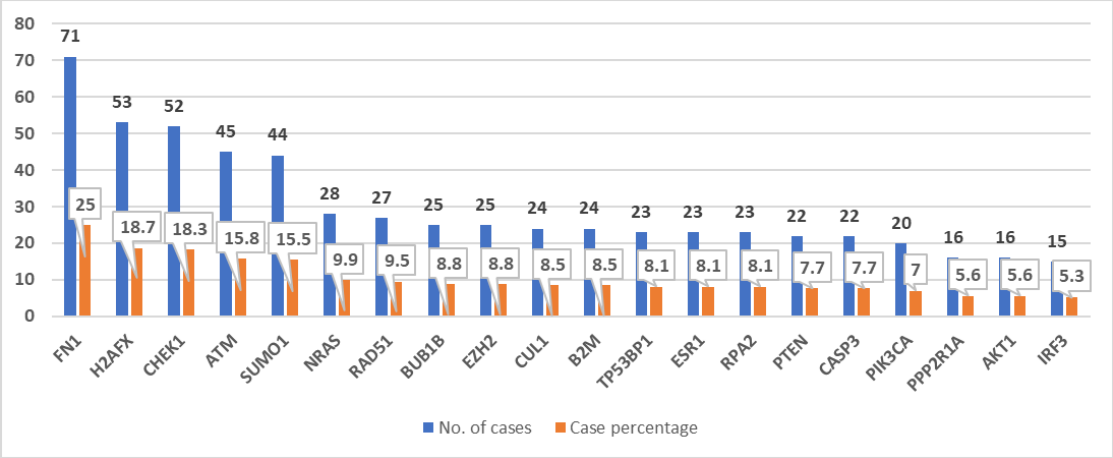


Figure 31. Target genes with significant copy number loss in CESC

Likewise, from 529 HNSCC samples, 524 cases have mutations (503 cases) or CNV data (515 cases) for undertaken target genes. OncoGrid of these alterations from HNSCC samples represents a complete landscape of 50 frequently mutated genes along with CNVs (loss and gain) (**Figure 32**). The top 20 most frequently mutated genes are illustrated in **Figure 33**. Frequently mutated genes among all are TP53 (17p13.1), NOTCH1 (9q34.3), PIK3CA (3q26.32), EP300 (22q13.2), CREBBP (16p13.3), HRAS (11p15.5), TLR4 (9q33.1), EGFR (7p11.2), FN1 (2q35), ATM (11q22.3), TP53BP1 (15q15.3), PTEN (10q23.31), POLR2A (17p13.1), SMAD4 (18q21.2), BRCA1 (17q21.31), EGF (4q25), HGF (7q21.11), ERBB2 (17q12), STAT1 (2q32.2), and PTK2 (8q24.3) (**Figure 33**). Overall, 1092 mutations are present in target genes. Most mutated genes among all are TP53, NOTCH1, CREBBP, PIK3CA, EP300, TLR4, EGFR, FN1, TP53BP1, ATM (**Figure 34**). The number of mutations in most mutated genes is presented in **Figure 34**. Similarly, **Figures 35** and **36** depict high CNV (gain and loss) genes, respectively.

Additionally, we have also analyzed the cross-mutated target genes between CESC and HNSCC samples. Overall, 49 common mutations were found between both the carcinomas (CESC (103 samples), HNSCC (130) of 20 genes (**Figure 37 and Table 10**). The most frequently mutated genes are PIK3CA, TP53, MAPK1, EP300, PTEN, and ERBB2 (**Figure 37**).

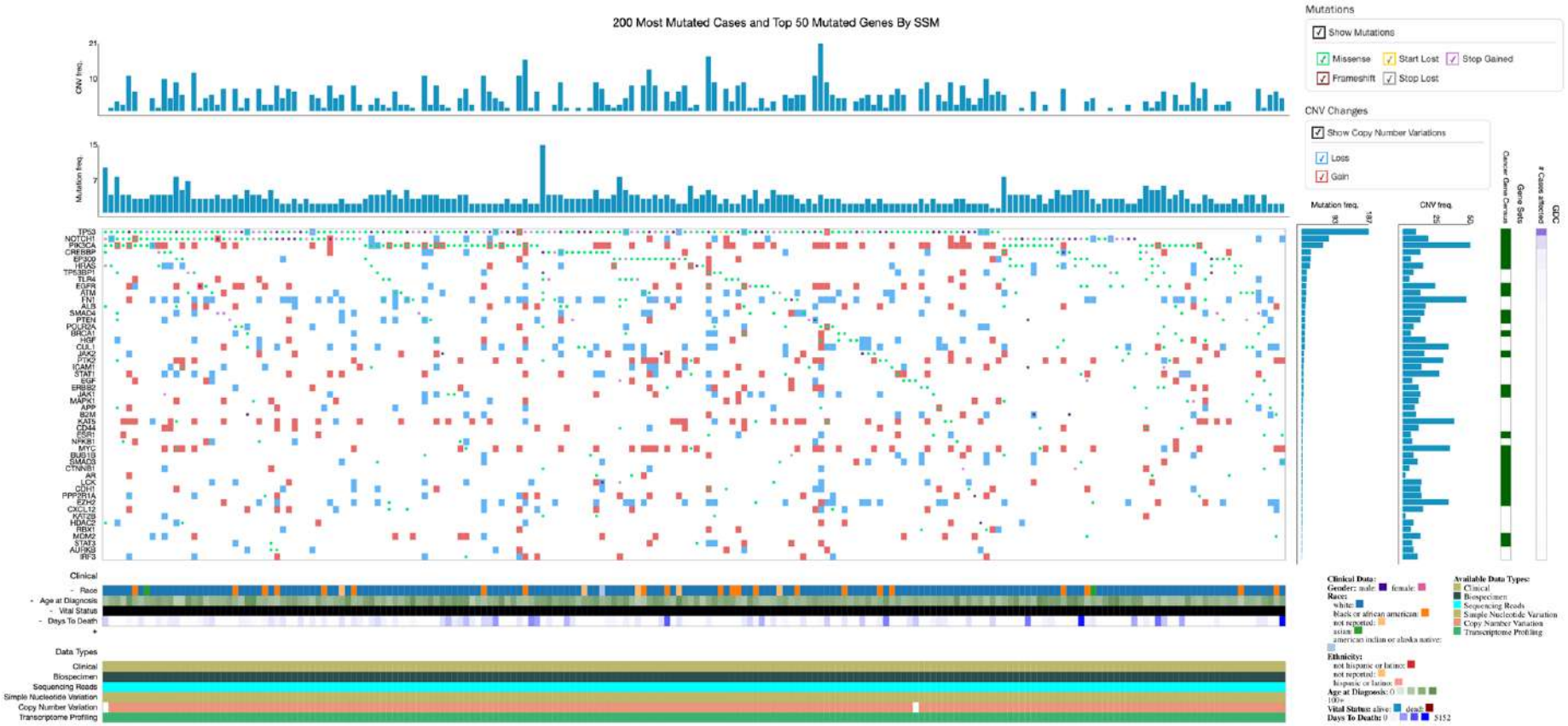


Figure 32. OncoGrid depicting top 50 genes with genomic alterations (mutations and CNVs) in HNSCC

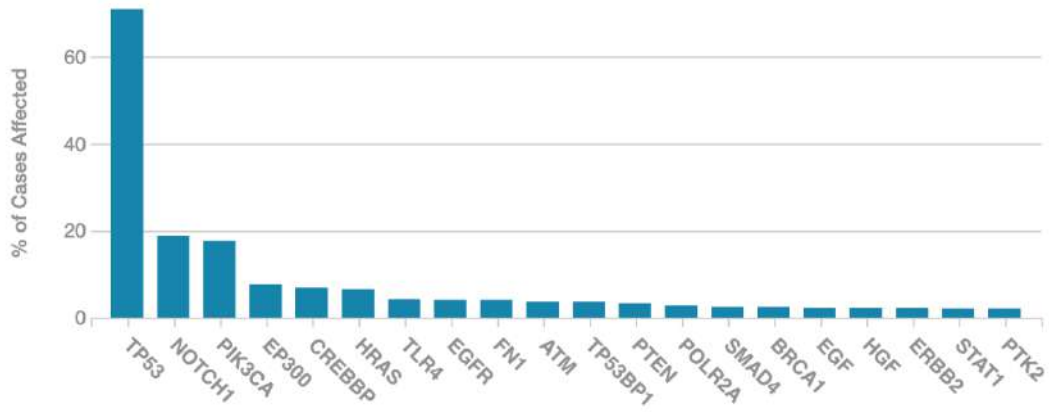


Figure 33. The most frequently mutated genes among targets in HNSCC samples

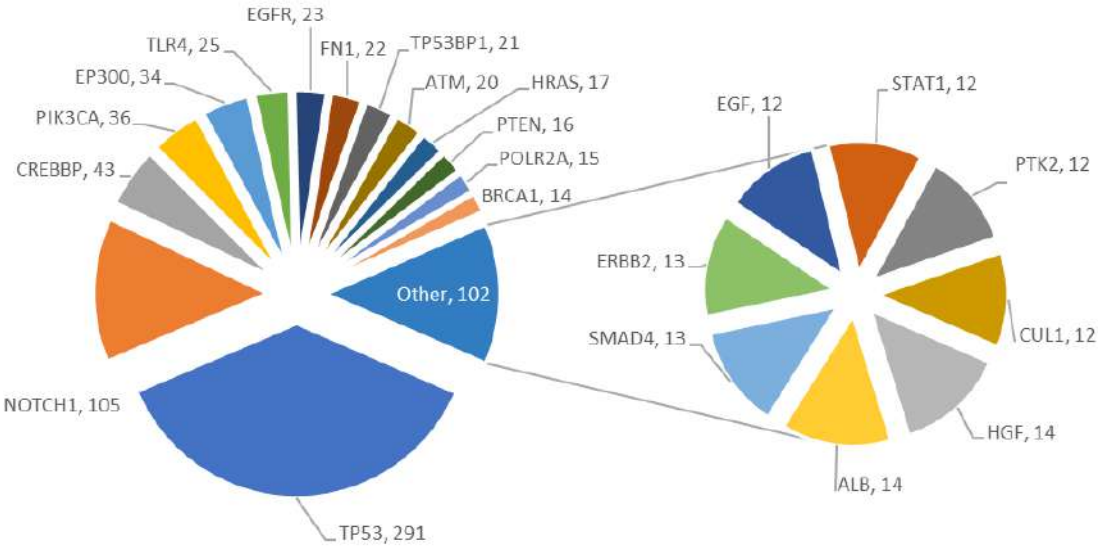


Figure 34. The most mutated target genes and number of mutations in HNSCC samples

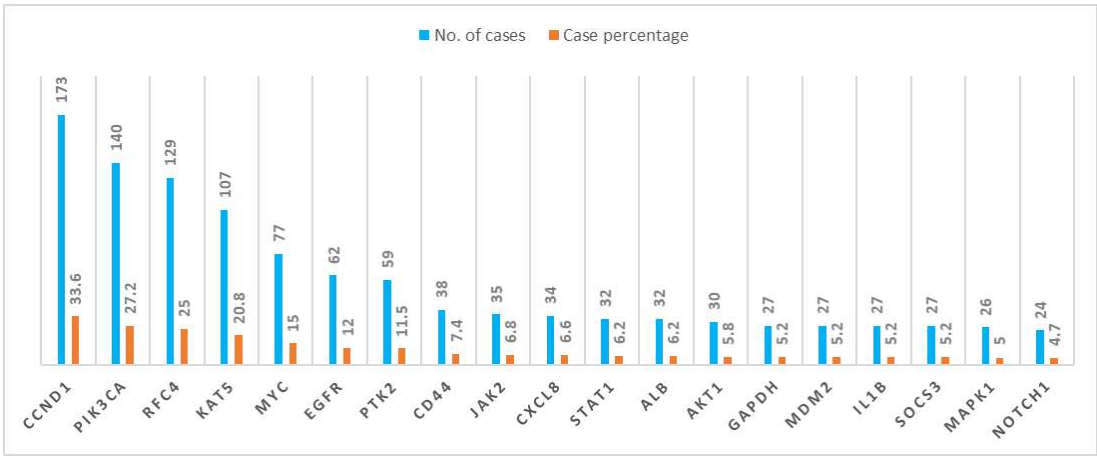


Figure 35. Genes with substantial copy number gain in HNSCC

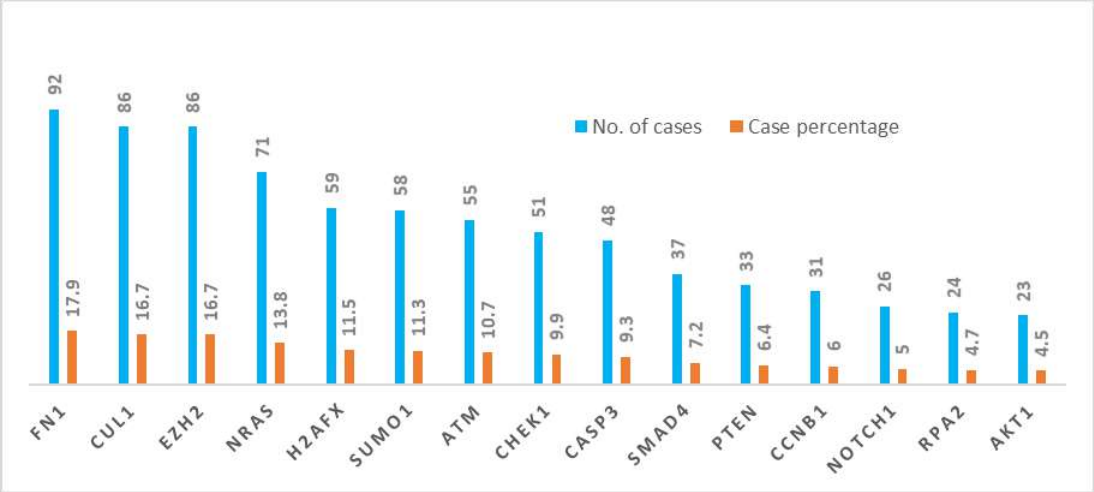


Figure 36. Genes with substantial copy number loss in HNSCC

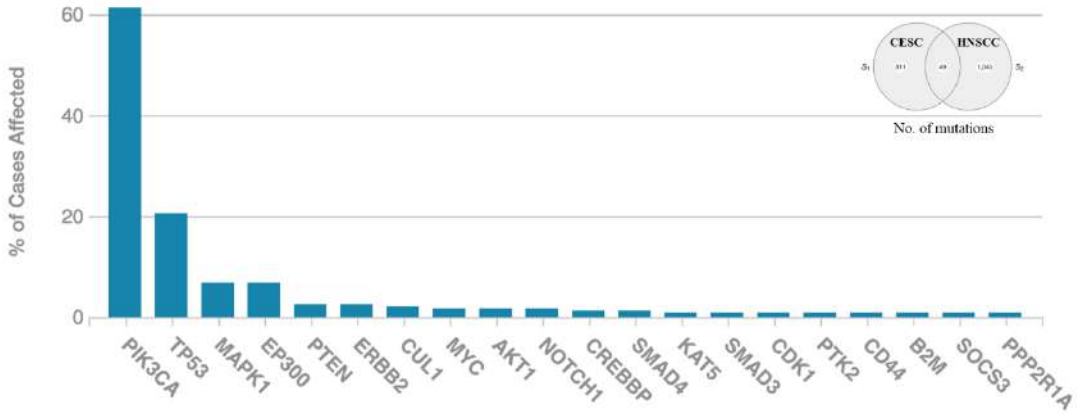


Figure 37. Frequently mutated genes in both CESC and HNSCC samples

Potential drugs and interactions between target genes using Open Targets Platform

Prioritized list of target genes was analyzed for the potential drugs targeting these proteins utilizing the Open target platform. Drug-target network is catalogued and presented (**Figure 38 and Table 11**). Overall, we identified 230 drugs pertaining to 41 target proteins. These drugs belong to different molecule types, i.e., Small molecules (161), Antibody (58), Protein (7), Enzyme (1), Oligosaccharide (1) and unknown (2) (**Figure 39A**). Among all, maximum drugs are targeting AR (26), ESR1 (26), EGFR (25), PIK3CA (17) and ERBB2 (14) (**Figure 39B**). Additionally, interactions between targets were also obtained utilizing the OmniPath DB, which provide comprehensive knowledge from 115 databases with focus on signaling pathways. This provides 707 enzyme-substrate relationship, 150 pathway-protein interactions and 355 protein-protein interactions between targets (**Figure 40**).

Table 10. List of 49 common mutations between CESC and HNSCC affecting 20 genes

DNA Change	Type	Consequences	# Affected Cases in Cohort	Impact
chr3:g.179218303G>A	Substitution	Missense PIK3CA E545K	64 / 233,27.47%	VEP: MODERATE, SIFT: deleterious - score 0.02, PolyPhen: probably_damaging - score 0.959
chr3:g.179218294G>A	Substitution	Missense PIK3CA E542K	42 / 233,18.03%	VEP: MODERATE, SIFT: deleterious - score 0.04, PolyPhen: probably_damaging - score 0.96
chr22:g.21772875C>T	Substitution	Missense MAPK1 E322K	16 / 233,6.87%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.999
chr3:g.179234297A>G	Substitution	Missense PIK3CA H1047R	15 / 233,6.44%	VEP: MODERATE, SIFT: tolerated - score 0.11, PolyPhen: possibly_damaging - score 0.529
chr17:g.7673776G>A	Substitution	Missense TP53 R282W	11 / 233,4.72%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.997
chr22:g.41169525G>A	Substitution	Missense EP300 D1399N	9 / 233,3.86%	VEP: MODERATE, PolyPhen: probably_damaging - score 1
chr17:g.7674945G>A	Substitution	Stop Gained TP53 R196*	8 / 233,3.43%	VEP: HIGH
chr3:g.179221146G>A	Substitution	Missense PIK3CA E726K	8 / 233,3.43%	VEP: MODERATE, SIFT: tolerated - score 0.36, PolyPhen: benign - score 0.4
chr17:g.7673788G>A	Substitution	Missense TP53 P278S	6 / 233,2.58%	VEP: MODERATE, SIFT: deleterious - score 0.03, PolyPhen: probably_damaging - score 1

DNA Change	Type	Consequences	# Affected Cases in Cohort	Impact
chr17:g.39711955C>T	Substitution	Missense ERBB2 S310F	6 / 233,2.58%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.994
chr17:g.7673767C>T	Substitution	Missense TP53 E285K	5 / 233,2.15%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.985
chr8:g.127738699C>T	Substitution	Missense MYC S161L	4 / 233,1.72%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: possibly_damaging - score 0.888
chr17:g.7674957G>A	Substitution	Stop Gained TP53 Q192*	4 / 233,1.72%	VEP: HIGH
chr14:g.104780214C>T	Substitution	Missense AKT1 E17K	4 / 233,1.72%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.999
chr3:g.179210192T>C	Substitution	Missense PIK3CA C420R	3 / 233,1.29%	VEP: MODERATE, SIFT: tolerated - score 0.21, PolyPhen: possibly_damaging - score 0.893
chr7:g.148787094G>A	Substitution	Missense CUL1 E485K	3 / 233,1.29%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.991
chr16:g.3738617G>A	Substitution	Missense CREBBP R1446C	3 / 233,1.29%	VEP: MODERATE, PolyPhen: unknown - score 0
chr10:g.87933147C>T	Substitution	Stop Gained PTEN R130*	3 / 233,1.29%	VEP: HIGH
chr3:g.179210291G>A	Substitution	Missense PIK3CA E453K	3 / 233,1.29%	VEP: MODERATE, SIFT: tolerated - score 0.06, PolyPhen: possibly_damaging - score 0.801

DNA Change	Type	Consequences	# Affected Cases in Cohort	Impact
chr10:g.87933148G>A	Substitution	Missense PTEN R130Q	3 / 233,1.29%	VEP: MODERATE, SIFT: deleterious - score 0.02, PolyPhen: probably_damaging - score 0.989
chr18:g.51065549G>A	Substitution	Missense SMAD4 R361H	3 / 233,1.29%	VEP: MODERATE, SIFT: deleterious - score 0.03, PolyPhen: probably_damaging - score 1
chr3:g.179218307A>G	Substitution	Missense PIK3CA Q546R	3 / 233,1.29%	VEP: MODERATE, SIFT: deleterious - score 0.02, PolyPhen: probably_damaging - score 0.984
chr22:g.41172586G>A	Substitution	Missense EP300 E1514K	3 / 233,1.29%	VEP: MODERATE, PolyPhen: probably_damaging - score 0.995
chr17:g.7673537G>A	Substitution	Stop Gained TP53 Q331*	3 / 233,1.29%	VEP: HIGH
chr17:g.7674250C>A	Substitution	Missense TP53 C238F	3 / 233,1.29%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 1
chr3:g.179199088G>A	Substitution	Missense PIK3CA R88Q	3 / 233,1.29%	VEP: MODERATE, SIFT: tolerated - score 0.06, PolyPhen: probably_damaging - score 0.998
chr3:g.179234176G>C	Substitution	Missense PIK3CA G1007R	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.1, PolyPhen: possibly_damaging - score 0.817
chr7:g.148786567G>A	Substitution	Missense CUL1 E439K	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.946
chr22:g.41160723G>A	Substitution	Splice Donor EP300 X1224_splice	2 / 233,0.86%	VEP: HIGH

DNA Change	Type	Consequences	# Affected Cases in Cohort	Impact
chr9:g.136518232C>T	Substitution	Missense NOTCH1 C387Y	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.999
chr17:g.7673781C>T	Substitution	Missense TP53 R280K	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0.04, PolyPhen: possibly_damaging - score 0.83
chr3:g.179203761T>G	Substitution	Missense PIK3CA V344G	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.987
chr3:g.179234297A>T	Substitution	Missense PIK3CA H1047L	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.44, PolyPhen: benign - score 0.085
chr9:g.136516000G>T	Substitution	Stop Gained NOTCH1 Y550*	2 / 233,0.86%	VEP: HIGH
chr17:g.7673781C>G	Substitution	Missense TP53 R280T	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.947
chr17:g.78358641G>A	Substitution	Missense SOCS3 S152F	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.11, PolyPhen: benign - score 0.003
chr3:g.179210186G>A	Substitution	Missense PIK3CA E418K	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.06, PolyPhen: possibly_damaging - score 0.886
chr15:g.44711583delCT	Deletion	Frameshift B2M L15Ffs*41	2 / 233,0.86%	VEP: HIGH
chr19:g.52226009C>T	Substitution	3 Prime UTR PPP2R1A	2 / 233,0.86%	VEP: MODIFIER

DNA Change	Type	Consequences	# Affected Cases in Cohort	Impact
chr8:g.140864347C>T	Substitution	Missense PTK2 E139K	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.46, PolyPhen: benign - score 0.048
chr11:g.65713787G>A	Substitution	Missense KAT5 R210H	2 / 233,0.86%	VEP: MODERATE, SIFT: tolerated - score 0.13, PolyPhen: benign - score 0.022
chr17:g.7673809C>T	Substitution	Missense TP53 E271K	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 0.999
chr11:g.35206118C>T	Substitution	Missense CD44 S430L	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0.02, PolyPhen: benign - score 0.164
chr17:g.7675216C>G	Substitution	Missense TP53 K132N	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: probably_damaging - score 1
chr17:g.7675136G>A	Substitution	Missense TP53 A159V	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: possibly_damaging - score 0.9
chr10:g.60794092C>T	Substitution	3 Prime UTR CDK1	2 / 233,0.86%	VEP: MODIFIER
chr3:g.179199048C>G	Substitution	Missense PIK3CA Q75E	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0.01, PolyPhen: possibly_damaging - score 0.872
chr15:g.67190505C>T	Substitution	Missense SMAD3 S416F	2 / 233,0.86%	VEP: MODERATE, SIFT: deleterious - score 0, PolyPhen: possibly_damaging - score 0.807
chr22:g.41170517G>C	Substitution	Missense EP300 W1466C	2 / 233,0.86%	VEP: MODERATE, PolyPhen: probably_damaging - score 1

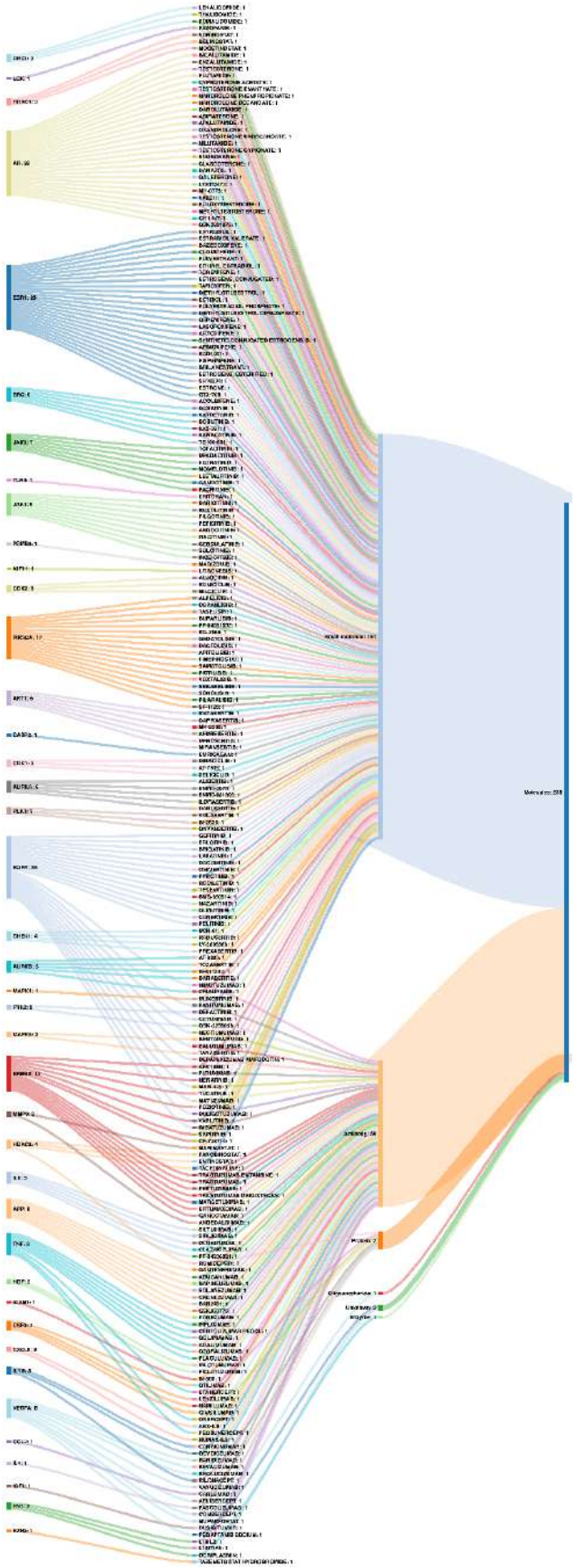


Figure 38. Sankey plot depicting targets, drugs and drug categories

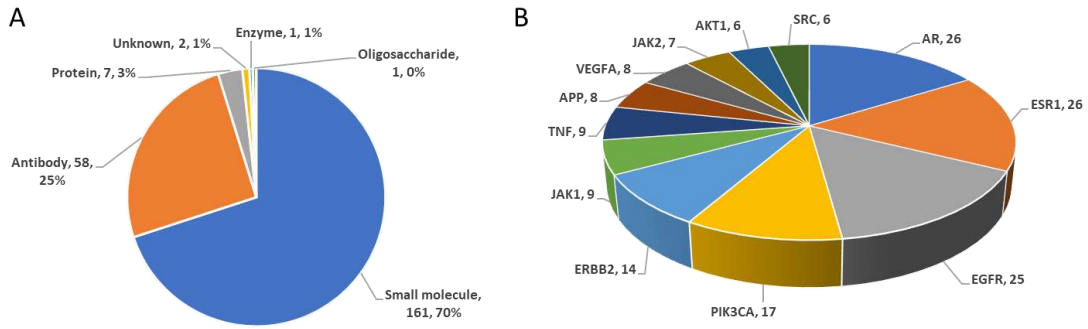


Figure 39. (A) Distribution of drug types, (B) Proteins with maximum targeting drugs

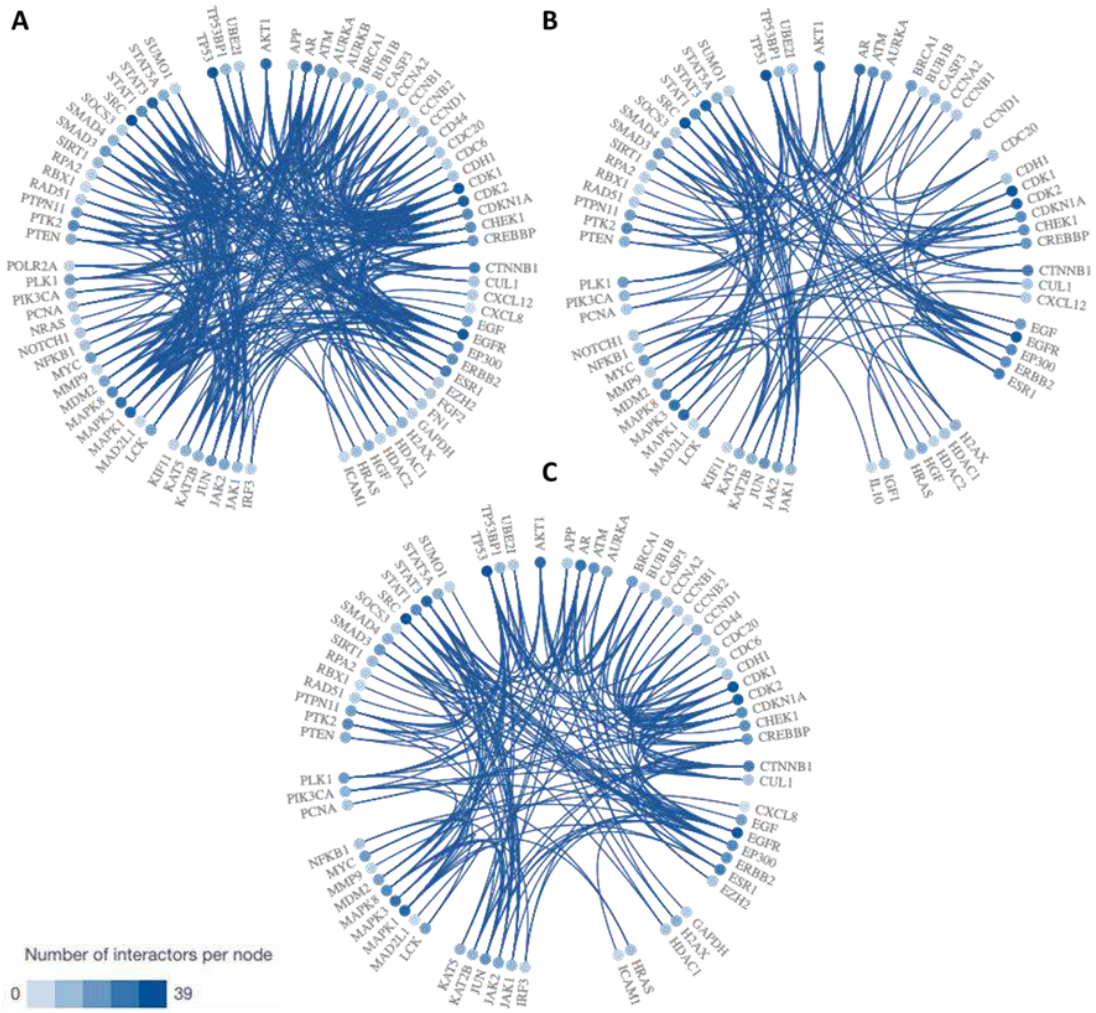


Figure 40. Distinct protein interactions. (A) Enzyme-substrate (707), (B) Pathway (150), (C) PPI (355 int)

Table 11. Potential drugs pertaining to different targets

Drug	*Target*	*Disease*	*Molecule type*
BICALUTAMIDE	AR	prostate cancer	SM
DASATINIB	SRC	acute lymphoblastic leukemia	SM
LENALIDOMIDE	RBX1	multiple myeloma	SM
AFLIBERCEPT	VEGFA	wet macular degeneration	Protein
ENZALUTAMIDE	AR	neoplasm	SM
ESTRADIOL	ESR1	hypogonadism	SM
TESTOSTERONE	AR	Hypogonadotropic hypogonadism	SM
INFLIXIMAB	TNF	rheumatoid arthritis	Antibody
RANIBIZUMAB	VEGFA	retinopathy	Antibody
GEFITINIB	EGFR	non-small cell lung carcinoma	SM
ETANERCEPT	TNF	rheumatoid arthritis	Protein
ALPELISIB	PIK3CA	neoplasm	SM
TRASTUZUMAB EMTANSINE	ERBB2	neoplasm	Antibody
ERLOTINIB	EGFR	non-small cell lung carcinoma	SM
THALIDOMIDE	RBX1	Mantle cell lymphoma	SM
ESTRADIOL VALERATE	ESR1	infertility	SM
PANOBINOSTAT	HDAC2	primary myelofibrosis	SM
COPANLISIB	PIK3CA	neoplasm	SM
BEVACIZUMAB	VEGFA	non-small cell lung carcinoma	Antibody
TOFACITINIB	JAK2	Takayasu arteritis	SM
CERTOLIZUMAB PEGOL	TNF	immune system disease	Antibody
VANDETANIB	SRC	thyroid cancer	SM
BRIGATINIB	EGFR	neoplasm	SM
BARICITINIB	JAK1	rheumatoid arthritis	SM
BAZEDOXIFENE	ESR1	obesity	SM
ROMIDEPSIN	HDAC2	neoplasm	Protein
CLOMIPHENE	ESR1	anovulation	SM
TRASTUZUMAB	ERBB2	breast carcinoma	Antibody
FULVESTRANT	ESR1	breast carcinoma	SM
AFATINIB	ERBB2	non-small cell lung carcinoma	SM
ETHINYL ESTRADIOL	ESR1	infertility	SM
TOREMIFENE	ESR1	breast carcinoma	SM
GOLIMUMAB	TNF	immune system disease	Antibody
LAPATINIB	EGFR	cancer	SM
PERTUZUMAB	ERBB2	neoplasm	Antibody
VORINOSTAT	HDAC1	neoplasm	SM

Drug	Target	Disease	Molecule type
ADALIMUMAB	TNF	ulcerative colitis	Antibody
NIMOTUZUMAB	EGFR	pancreatic carcinoma	Antibody
PAZOPANIB	LCK	neoplasm	SM
NERATINIB	ERBB2	neoplasm	SM
UPADACITINIB	JAK2	rheumatoid arthritis	SM
RUXOLITINIB	JAK1	polycythemia vera	SM
ESTROGENS, CONJUGATED	ESR1	postmenopausal osteoporosis	SM
PEGAPTANIB SODIUM	VEGFA	retinal vein occlusion	Unknown
TAMOXIFEN	ESR1	breast cancer	SM
FLUTAMIDE	AR	prostate adenocarcinoma	SM
CANAKINUMAB	IL1B	Familial Mediterranean fever	Antibody
CYPROTERONE ACETATE	AR	polycystic ovary syndrome	SM
DACOMITINIB	EGFR	neoplasm	SM
TESTOSTERONE ENANTHATE	AR	hypogonadotropic hypogonadism	SM
TAZEMETOSTAT HYDROBROMIDE	EZH2	sarcoma	Unknown
PANITUMUMAB	EGFR	neoplasm	Antibody
CETUXIMAB	EGFR	metastatic colorectal cancer	Antibody
DIETHYLSTILBESTROL	ESR1	neoplasm	SM
FEDRATINIB	JAK2	neoplasm	SM
NANDROLONE PHENPROPIONATE	AR	eye disease	SM
NECITUMUMAB	EGFR	non-small cell lung carcinoma	Antibody
BELINOSTAT	HDAC1	unspecified peripheral T-cell lymphoma	SM
NANDROLONE DECANOATE	AR	eye disease	SM
RILONACEPT	IL1B	Muckle-Wells syndrome	Protein
POMALIDOMIDE	RBX1	immune system disease	SM
DAROLUTAMIDE	AR	neoplasm	SM
CONBERCEPT	VEGFA	retinal detachment	Protein
ABIRATERONE	AR	prostate cancer	SM
TUCATINIB	ERBB2	HER2 Positive Breast Carcinoma	SM
APALUTAMIDE	AR	neoplasm	SM
OXANDROLONE	AR	HIV wasting syndrome	SM
ESTRIOL	ESR1	urinary tract infection	SM
SILTUXIMAB	IL6	Giant Lymph Node Hyperplasia	Antibody

Drug	Target	Disease	Molecule type
BOSUTINIB	SRC	neoplasm	SM
POLYESTRADIOL PHOSPHATE	ESR1	neoplasm	SM
TESTOSTERONE UNDECANOATE	AR	erectile dysfunction	SM
OSIMERTINIB	EGFR	non-small cell lung carcinoma	SM
DIETHYLSTILBESTROL DIPHOSPHATE	ESR1	neoplasm	SM
NILUTAMIDE	AR	neoplasm	SM
OCRIPLASMIN	FN1	eye disease	Enzyme
TESTOSTERONE CYPIONATE	AR	Klinefelter's syndrome	SM
OSPEMIFENE	ESR1	sexual dysfunction	SM
BROLUCIZUMAB	VEGFA	ocular vascular disease	Antibody
OTILIMAB	CSF2	rheumatoid arthritis	Antibody
GEVOKIZUMAB	IL1B	uveitis	Antibody
ENOBOSARM	AR	non-small cell lung carcinoma	SM
FILGOTINIB	JAK1	Crohn's disease	SM
SIRUKUMAB	IL6	rheumatoid arthritis	Antibody
IPATASERTIB	AKT1	triple-negative breast cancer	SM
GANTENERUMAB	APP	Alzheimer's disease	Antibody
PEFICITINIB	JAK1	rheumatoid arthritis	SM
PYROTINIB	EGFR	breast cancer	SM
ONERCEPT	TNF	psoriatic arthritis	Protein
ENTINOSTAT	HDAC2	breast cancer	SM
MARIMASTAT	MMP9	lung cancer	SM
ANDECALIXIMAB	MMP9	gastric adenocarcinoma	Antibody
ADUCANUMAB	APP	Alzheimer's disease	Antibody
BAPINEUZUMAB	APP	Alzheimer's disease	Antibody
ABROCITINIB	JAK1	atopic eczema	SM
ZALUTUMUMAB	EGFR	head and neck malignant neoplasia	Antibody
MOMELOTINIB	JAK2	pancreatic ductal adenocarcinoma	SM
CLASCOTERONE	AR	acne	SM
KX2-391	SRC	actinic keratosis	SM
DEPATUXIZUMAB MAFODOTIN	EGFR	glioblastoma multiforme	Antibody
ROCILETINIB	EGFR	non-small cell lung carcinoma	SM
LASOFOXIFENE	ESR1	osteoporosis	SM
L19IL2	FN1	melanoma	Antibody

Drug	Target	Disease	Molecule type
CAPIVASERTIB	AKT1	breast cancer	SM
SOLANEZUMAB	APP	Alzheimer's disease	Antibody
VOLASERTIB	PLK1	acute myeloid leukemia	SM
TASELISIB	PIK3CA	breast cancer	SM
ARZOXIFENE	ESR1	postmenopausal osteoporosis	SM
ERITORAN	TLR4	Sepsis	SM
OLOKIZUMAB	IL6	rheumatoid arthritis	Antibody
DINACICLIB	CDK1	chronic lymphocytic leukemia	SM
CRENEZUMAB	APP	Alzheimer's disease	Antibody
BUPARLISIB	PIK3CA	breast cancer	SM
MUPARFOSTAT	VEGFA	hepatocellular carcinoma	Oligosaccharide
TESEVATINIB	EGFR	non-small cell lung carcinoma	SM
LENZILUMAB	CSF2	pneumonia	Antibody
RILOTUMUMAB	HGF	gastric cancer	Antibody
FUTUXIMAB	EGFR	metastatic colorectal cancer	Antibody
ITACITINIB	JAK1	graft versus host disease	SM
DANAZOL	AR	infertility	SM
BAN2401	APP	Alzheimer's disease	Antibody
TRASTUZUMAB DERUXTECAN	ERBB2	breast cancer	Antibody
ALISERTIB	AURKA	unspecified peripheral T-cell lymphoma	SM
TACEDINALINE	HDAC2	lung cancer	SM
GALETERONE	AR	prostate cancer	SM
MARIZOMIB	PSMB8	glioblastoma multiforme	SM
LESTAURTINIB	JAK2	childhood T acute lymphoblastic leukemia	SM
SYNTHETIC CONJUGATED ESTROGENS, B	ESR1	menopause	SM
LY2452473	AR	erectile dysfunction	SM
BMS-690514	EGFR	breast cancer	SM
LITRONESIB	KIF11	small cell lung carcinoma	SM
ALVOCIDIB	CDK2	endometrial cancer	SM
RONICICLIB	CDK2	small cell lung carcinoma	SM
AFIMOXIFENE	ESR1	breast ductal carcinoma in situ	SM
MAB-425	EGFR	central nervous system cancer	Antibody

Drug	Target	Disease	Molecule type
MOCETINOSTAT	HDAC1	chronic lymphocytic leukemia	SM
RAD1901	ESR1	Hot flashes	SM
POZIOTINIB	ERBB2	non-small cell lung carcinoma	SM
VARLITINIB	ERBB2	cholangiocarcinoma	SM
AT-9283	AURKB	multiple myeloma	SM
SAPITINIB	ERBB2	breast cancer	SM
NAZARTINIB	EGFR	lung cancer	SM
UCN-01	CHEK1	fallopian tube cancer	SM
FISPEMIFENE	ESR1	hypogonadism	SM
GANDOTINIB	JAK2	neoplasm	SM
EMRICASAN	CASP3	cirrhosis of liver	SM
BI-2536	PLK1	acute myeloid leukemia	SM
TOZASERTIB	AURKB	leukemia	SM
PF-04691502	PIK3CA	breast neoplasm	SM
RABUSERTIB	CHEK1	non-small cell lung carcinoma	SM
MK-2206	AKT1	hepatocellular carcinoma	SM
RG-7666	PIK3CA	glioblastoma multiforme	SM
SARACATINIB	SRC	pancreatic adenocarcinoma	SM
PEGSUNERCEPT	TNF	rheumatoid arthritis	Protein
GEDATOLISIB	PIK3CA	endometrial neoplasm	SM
MILCICLIB	CDK2	hepatocellular carcinoma	SM
BI-505	ICAM1	multiple myeloma	Antibody
AFURESERTIB	AKT1	Langerhans Cell Histiocytosis	SM
DACTOLISIB	PIK3CA	neoplasm	SM
UPROSERTIB	AKT1	breast carcinoma	SM
ENMD-2076	AURKA	clear cell adenocarcinoma	SM
MATUZUMAB	EGFR	non-small cell lung carcinoma	Antibody
CLAZAKIZUMAB	IL6	rheumatoid arthritis	Antibody
AT-7519	CDK1	chronic lymphocytic leukemia	SM
ABX-IL8	CXCL8	chronic bronchitis	Antibody
MARGETUXIMAB	ERBB2	breast cancer	Antibody
ENMD-981693	AURKA	breast cancer	SM
APITOLISIB	PIK3CA	endometrial carcinoma	SM
LY-2606368	CHEK1	cancer	SM
FICLATUZUMAB	HGF	acute myeloid leukemia	Antibody
ILORASERTIB	AURKA	cancer	SM
GSK933776	APP	atrophic macular degeneration	Antibody

Drug	Target	Disease	Molecule type
SELICICLIB	CDK1	pituitary-dependent Cushing's disease	SM
PF-04236921	IL6	Crohn's disease	Antibody
MK-0773	AR	sarcopenia	SM
FIMEPINOSTAT	PIK3CA	thyroid cancer	SM
OZORALIZUMAB	TNF	rheumatoid arthritis	Antibody
PLACULUMAB	TNF	rheumatoid arthritis	Antibody
PONEZUMAB	APP	cerebral amyloid angiopathy	Antibody
NAMILUMAB	CSF2	rheumatoid arthritis	Antibody
VK5211	AR	hip fracture	SM
ERTUMAXOMAB	ERBB2	breast cancer	Antibody
BRILANESTRANT	ESR1	breast cancer	SM
ESTROGENS, ESTERIFIED	ESR1	breast cancer	SM
SAMOTOLISIB	PIK3CA	non-small cell lung carcinoma	SM
DANUSERTIB	AURKA	multiple myeloma	SM
PICTILISIB	PIK3CA	breast cancer	SM
OLMUTINIB	EGFR	non-small cell lung carcinoma	SM
PACRITINIB	JAK2	myeloproliferative disorder	SM
VOXTALISIB	PIK3CA	ovarian cancer	SM
DULIGOTUZUMAB	EGFR	colorectal carcinoma	Antibody
DEFACTINIB	PTK2	pancreatic ductal adenocarcinoma	SM
CARLUMAB	CCL2	pulmonary fibrosis	Antibody
CANERTINIB	EGFR	lung neoplasm	SM
FLUOXYMESTERONE	AR	breast cancer	SM
ULIXERTINIB	MAPK1	Uveal Melanoma	SM
BI-811283	AURKB	acute myeloid leukemia	SM
METHYLTESTOSTERONE	AR	menopause	SM
SR16234	ESR1	breast cancer	SM
CERDULATINIB	JAK1	Vitiligo	SM
PREXASERTIB	CHEK1	small cell lung carcinoma	SM
CP-724714	ERBB2	metastasis	SM
ESTRONE	ESR1	obesity	SM
CR 1447	AR	breast adenocarcinoma	SM
GTX-758	ESR1	prostate cancer	SM
IMGATUZUMAB	EGFR	colorectal carcinoma	Antibody
SERABELISIB	PIK3CA	endometrial neoplasm	SM
SONOLISIB	PIK3CA	glioblastoma multiforme	SM
BARASERTIB	AURKB	acute myeloid leukemia	SM

Drug	Target	Disease	Molecule type
GANCOTAMAB	ERBB2	breast cancer	Antibody
BENTAMAPIMOD	MAPK8	endometriosis	SM
PILARALISIB	PIK3CA	endometrial neoplasm	SM
GSK-2256098	PTK2	pancreatic adenocarcinoma	SM
CHIAURANIB	AURKB	ovarian cancer	SM
GSK2881078	AR	Cachexia	SM
L19TNFA	FN1	soft tissue sarcoma	Antibody
PELITINIB	EGFR	colonic neoplasm	SM
VANUCIZUMAB	VEGFA	colorectal carcinoma	Antibody
TG100-801	SRC	wet macular degeneration	SM
SF-1126	PIK3CA	head and neck squamous cell carcinoma	SM
PASCOLIZUMAB	IL4	Tuberculosis	Antibody
SOLCITINIB	JAK1	systemic lupus erythematosus	SM
ACOLBIFENE	ESR1	breast cancer	SM
TANZISERTIB	MAPK8	lupus erythematosus	SM
DUSIGITUMAB	IGF1	breast cancer	Antibody
INCB-047986	JAK1	rheumatoid arthritis	SM
MIRANSERTIB	AKT1	Proteus syndrome	SM
HUMAX-IL8	CXCL8	COVID-19	Antibody
ONVANSERTIB	PLK1	prostate cancer	SM
GIMSILUMAB	CSF2	acute respiratory distress syndrome	Antibody

SM: Small molecule

Discussion

HPV is a pathogenic and infectious oncovirus causative agent of various cancers globally that pose a global burden of mortality (Arbyn et al., 2020; Bray et al., 2018; de Martel et al., 2020). One of the most important consequences in HPV infection and carcinogenesis is integration events that lead to the alteration of host genes (Hu et al., 2015; Kumar Gupta and Kumar, 2015; Rusan et al., 2015). This was reported to contribute towards the aberrant proliferation, miRNA dysregulation, genomic instability, genomic structural alterations, cellular immortalization, epigenetic alterations and malignant progression (Hu et al., 2015; Kumar Gupta and Kumar, 2015; Peter et al., 2010). Disruption can also boost oncoprotein expression and leads to the loss of function of cell cycle checkpoints, DNA repair mechanisms and tumor suppressor genes (Hu et al., 2015; McBride and Warburton, 2017; Rusan et al., 2015).

OncoHPV-PPI-omics provide an illustrative approach combining network systems biology and multi-omics analysis towards the identification of key and core therapeutics targets, biological processes, enriched pathways, genomic alterations, and potential known drugs with respect to the HPV pathogenesis including cervical squamous cell carcinoma (CSCC) and head and neck squamous cell carcinoma (HNSCC). Here, we analyzed the 1887 candidate genes excluding 96 duplicates combining 1520 HPV infection associated genes from "Open Targets Platform" (Carvalho-Silva et al., 2019) and 463 disrupted genes due to HPV integration events from HPVbase (Kumar Gupta and Kumar, 2015).

We integrate and consider different strategies. First, 1887 candidates were subjected to the protein-protein interactions network analysis with a high confidence score of 0.7. Further, the hub genes were identified to elucidate potential targets. Complete interactome consist of 1879 nodes and 20735 edges is analyzed based on four different methods, i.e., Degree, EPC, MNC and EcCentricity and the top 100 targets from each were identified (**Figure 20-23**). Common targets from each method were analyzed and identified as the PPI-network core genes (**Table 7 and Figure 24**). Further, the top 100 potential targets with the highest degree and correlation are marked for the potential targets in the HPV infection and pathogenesis (**Figure 20**). Out of these genes, 70 targets were classified in the Oncogenes, TFs, Kinases, Cytokines and growth factors, Translocated cancer genes, Tumor suppressor genes, and CD markers (**Figure 25**).

More, GO and pathway enrichment analysis was performed, which further substantiated our approach (**Figure 26**). We have identified binding (enzyme, receptor, transcription factor, kinase, protein, DNA) as the most prominent molecular functions (**Figure 26**), which represent critical factors towards HPV infection, integration and cell cycle alterations. Likewise, the major biological process represents regulation of the metabolic process, signaling, cell cycle, proliferation, and apoptosis (cell death). Moreover, major pathways were also illuminated in the analysis. Enriched pathway mainly represents Pathways in cancer (including Endometrial, Colorectal, Breast, Bladder, Prostate), viral carcinogenicity, Human papillomavirus infection, EGFR tyrosine kinase inhibitor resistance, PI3K-Akt signaling pathway, JAK-STAT signaling pathway, ErbB signaling pathway, Cell

cycle and MicroRNAs in cancer (**Figure 26**). Our findings were further corroborated through gene set enrichment analysis (GSEA) to explore different hallmarks. Distinct significant functional hallmarks, i.e., G2M checkpoint, E2F-Targets, Apoptosis, Allograft rejection, TNFA signaling via NFKB, PI3K_AKT_MTOR signaling, IL6_JAK_STAT3 signaling, INTERFERON_GAMMA response, Inflammatory response, WNT_BETA_CATENIN signaling, MYC_TARGETS_V1, Apical junction, Epithelial-mesenchymal transition (EMT), Estrogen response late, P53 pathway, DNA repair, Spermatogenesis, and NOTCH signaling were identified (**Table 8**).

Simultaneously, significant regulatory miRNAs (**Table 9**) and human phenotypes were also identified. The potential target and noteworthy miRNAs are hsa-miR-155-5p, hsa-miR-34a-5p, hsa-miR-146a-5p, hsa-miR-203a-3p, hsa-miR-92a-3p, hsa-miR-193b-3p, hsa-miR-26a-5p, hsa-miR-145-5p, hsa-miR-24-3p, hsa-miR-199a-5p (**Table 9**). Similarly, as per expectations, phenotype ontologies are in-line and largely relevant to somatic mutation, neoplasm and carcinomas.

Therefore, our key and core target genes were further validated and prioritized utilizing the clinical data from CESC and HNSCC from TCGA hosted at GDC data portal and cBioPortal (2017; Briese et al., 2015; Pérez Sayáns et al., 2019; Weinstein et al., 2013). This includes the genomic alterations, i.e., mutations (mainly missense, and frameshift) and CNV (gain and loss). In the study, CESC cohort includes 293 samples representing clinical data for selected target genes from 307 samples. Likewise, HNSCC cohort of 524 cases has clinical data for target genes from 529 samples. OncoGrid represents a comprehensive landscape of significant and frequent mutations and CNVs in CESC (**Figure 27**) and HNSCC (**Figure 32**).

In the case of CESC, mainly missense mutation and specific CNVs gain and loss is identified. The most mutationally affected targets are PIK3CA, EP300, PTEN, CREBBP, NOTCH1, TP53, KRAS, ERBB2, TP53BP1, AR, SMAD4, MAPK1, POLR2A, EGFR, ATM, BRCA1, FN1, CUL1, and ESR1 (**Figure 28**). However, the higher number of mutations is present on EP300, PTEN, NOTCH1, CREBBP, PIK3CA, TP53, POLR2A, ERBB2, FN1, TP53BP1, etc. (**Figure 29**). Most importantly, certain genes such as PIK3CA (20.07% cases), RFC4 (19.01%), KAT5

(14.44%), MYC (13.03%), CCND1 (12.32%), PTK2 (11.62%), ERBB2 (8.10%), etc. were identified, which have significant copy number gains in CESC samples (**Figure 30**). Likewise, copy number loss of FN1 (25.00% cases), H2AFX (18.66%), CHEK1 (18.31%), ATM (15.85%), SUMO1 (15.49%), etc. are found noteworthy (**Figure 31**) and proposed to be relevant in HPV oncogenesis and could be used for drug and biomarker discovery.

In HNSCC, primarily missense, frameshift and stop gained mutations are dominant along with CNVs (loss and gain). The most affected targets among all are TP53, NOTCH1, PIK3CA, EP300, CREBBP, HRAS, TLR4, EGFR, FN1, ATM, TP53BP1, PTEN, POLR2A, SMAD4, BRCA1, EGF, HGF, ERBB2, STAT1 and PTK2 (**Figure 33**). Most mutations are found on the TP53, NOTCH1, CREBBP, PIK3CA, EP300, TLR4, EGFR, FN1, TP53BP1, ATM among all (**Figure 34**). Some target genes, i.e., CCND1 (33.59% cases), PIK3CA (27.18%), RFC4 (25.05%), KAT5 (20.78%), MYC (14.95%), EGFR (12.04%), PTK2 (11.46%) etc. have substantial copy number gains in HNSCC samples (**Figure 35**). Similarly, critical copy number loss of genes such as FN1 (17.86% cases), CUL1 (16.70%), EZH2 (16.70%), NRAS (13.79%), H2AFX (11.46%), SUMO1 (11.26%), ATM (10.68%), CHEK1 (9.90%) etc. are marked (**Figure 36**) and could be incorporated for potential therapeutic discovery.

Various high-throughput studies also advocated the use of distinct genomic, and transcriptomic applications and reported the crucial genes in CESC and HNSCC cases (2017; Ojesina et al., 2014; Rusan et al., 2015; Tuna and Amos, 2017). Like, Ojesina et. al. reported the somatic mutations in PIK3CA, PTEN, TP53, STK11, KRAS, MAPK1, EP300, FBXW7, NFE2L2, TP53, ERBB2 in CSCC and ELF3 and CBFB in adenocarcinoma (Ojesina et al., 2014). Another study by The Cancer Genome Atlas Research Network defines the APOBEC, SHKBP1, ERBB3, CASP8, HLA-A and TGFBR2 as novel mutated genes in 228 primary cervical cancers (2017). Likewise, Zhang et al. identify the differential expressed genes related to cervical intraepithelial neoplasia (Zhang et al., 2020) and report enrichment of E2F-Targets, G2M-Checkpoint, Mitotic-Spindle, and Spermatogenesis pathways. Another study performed a meta-analysis of transcriptomics data and revealed KAT2B, PCNA, CD86, PARP1, CDK1, GSK3B, WNK1, CRYAB, E2F4, ETS1, CUTL1, miRNAs (miR-192-5p, miR-193b-3p, and miR-215-5p) and some receptors like ephrin

(EPHA4, EPHA5), endothelin (EDNRA, EDNRB) and nuclear (NCOA3, NR2C1, NR2C2) as potential biomarkers and target in cervical cancer (Kori and Yalcin Arga, 2018). Moreover, various studies also provide a mutational landscape from HNSCC. Stransky et. al. analyzed whole-exome sequencing data and identified HNSCC mutated genes such as TP53, CDKN2A, PTEN, PIK3CA, HRAS, NOTCH1, IRF6, and TP63 (Stransky et al., 2011). Likewise, Seiwert et. al. report the comparative analysis between HPV+ and HPV- HNSCC (Seiwert et al., 2015). They show mutations in TP53, CDKN2A, MLL2, CUL3, NSD1, PIK3CA, and NOTCH in HPV- HNSCC. Also, mutation (FGFR2/3, DDX3X) and aberrations (PIK3CA, NOTCH1, KRAS, MLL2/3) were reported in HPV+ HNSCC (Seiwert et al., 2015). Likewise, Gaykalova et. al. defined genetic alterations in TP53, NOTCH1, FGFR3, CEBPA and FES (Gaykalova et al., 2014).

Furthermore, different studies also report the mis-regulation of different pathways mainly PI3K/Akt/mTOR signaling, Wnt/β-catenin/Notch, JAK/STAT Signaling and FGFR in HPV oncogenesis (Gupta et al., 2018; Morgan and Macdonald, 2020; Zhang et al., 2015; Zhang et al., 2016). Brand et. al. reported the role of HPV in the HER-3 associated PI3K signaling pathway in HPV+ HNSCC (Brand et al., 2017). Gaykalova et. al. describes the alteration in NOTCH signaling pathway (Gaykalova et al., 2014). Additionally, studies also demonstrate the role of FGFR2 and epithelial-mesenchymal transition (EMT) in HPV-related cancers (Ranieri et al., 2015; Zhang et al., 2016). More recently, Ren et. al. defined the activation of FGFR pathway in HPV positive cancer driven by HPV E2, E4 and E5 expression (Ren et al., 2020).

To aid, we have also explored the existing potential drugs targeting identified targets utilizing the Open Target Platform. Overall, 230 potential regimens targeting 41 proteins in different diseases were catalogued (**Figure 38 and Table 11**). These are mainly based on the small molecules, antibody and proteins (**Figure 39A**). Drug molecules are mainly targeting these protein targets, i.e., AR, ESR1, EGFR, PIK3CA, ERBB2, JAK1, TNF, APP, VEGFA, JAK2, AKT1, SRC, etc. with respect to distinct ailments (**Figure 39B)**. Furthermore, we also propose potential repurposing drug candidates like Dactolisib, Pilaralisib, Defactinib, Dacomitinib, Panitumumab, etc. These regimens could also be utilized for the drug repurposing in HPV infections (**Table 11)**.

Overall, we conclude and reveal known as well as novel significant core and key targets, TFs, miRNAs, pathways, functional hallmarks in HPV infection with significance from genomic alterations from clinical data and known potential drug-target relationships. It provides an understanding of genes, different mechanisms, pathways and biological functions contributing to the pathogenesis of HPVs and carcinogenicity. This may assist in formulating multi-dimensional strategies to prevent and treat HPV induced infections and carcinomas.

Development of human papillomavirus (HPV) genomic and therapeutic resource

Chapter 4. Development of human papillomavirus (HPV) genomic and therapeutic resource

Introduction

Human papillomaviruses (HPVs) are known to infect mucosal or cutaneous epithelial tissues. According to the malignant transformation competence, these are classified into distinct subgroups: high-risk HPVs (HR-HPVs highly carcinogenic) associated with diverse cancers and low-risk HPVs (LR-HPVs) which are linked mainly with genital warts. Persistence infection of HR-HPVs is extremely associated with cancer progression and can cause a diverse array of malignancies, i.e., cervical, oropharyngeal, penile, vulvar and anal carcinomas (Brianti et al., 2017; Crosbie et al., 2013; Munoz et al., 2003; zur Hausen, 2002). HR-HPV types usually 16 and 18 are prevalent in the etiology of human carcinomas and play a cardinal role in cervical cancer, which is the fourth most common cancer in women (de Martel et al., 2017; de Sanjose et al., 2010; Ferlay et al., 2015; Forman et al., 2012).

In the HPV carcinogenesis, HPV E6 and E7 oncoproteins are considered as the most preferred and ideal target for the therapeutic vaccines as they play a crucial part in the HPV mediated malignant transformations i.e. from low-grade cervical intraepithelial neoplasia (CIN 1) to high-grade CIN 2/3 and finally into invasive cervical cancer (ICC) (Hoppe-Seyler et al., 2018; Manzo-Merino et al., 2013; Mirabello et al., 2017). E6 and E7 protein degrade the p53 (apoptosis regulator) and the tumor suppressor retinoblastoma protein (pRb), respectively, which results in the disruption of cell apoptosis, cell life cycle regulation which leads to abnormal cell growth, host genomic instability and eventually cancer progression (Dadar et al., 2018; Doorbar et al., 2015; Mantovani and Banks, 2001; Moody and Laimins, 2010; zur Hausen, 2002). Along with this, E5 is also considered as an oncogene and several researchers also suggest their role in HPV carcinogenesis (Kim et al., 2010; Paolini et al., 2017).

HR- and LR-HPV types are most critical and bear a priority in terms of vaccine development against them. Several efforts are made to prevent HPV induced diseases by employing prophylactic and immunotherapeutic vaccine approaches (Chabeda et al., 2018; Dadar et al., 2018). Prophylactic vaccines based on the virus-like particles

(VLPs) from L1 capsid proteins were developed to resist HPV-induced malignancy (Harper, 2009; Lowy and Schiller, 2006). In view of this, earlier two HPV vaccines were developed to prevent HPV infection, a quadrivalent HPV-6/11/16/18 vaccine named as Merck's Gardasil® and a bivalent HPV-16/18 vaccine known as GlaxoSmithKline's Cervarix® (Harper et al., 2004; Siddiqui and Perry, 2006; Villa et al., 2005). Lately, a new vaccine named Gardasil®9 (human papillomavirus nonavalent vaccine, recombinant) was developed to protect against 9 types of HPVs (i.e. 6, 11, 16, 18, 31, 33, 45, 52 and 58) (Huh et al., 2017; Joura et al., 2015). However, these vaccines are ineffective at eliminating established infections (Hancock et al., 2018; Hildesheim et al., 2007; Hu and Ma, 2018; Hung et al., 2008). The viral capsid based vaccines are not able to affect infected basal cells due to the late expression of HPV capsid proteins in the replication cycle. Moreover, it failed to clear infection due to the non-productive infections of HPV associated cancers that lead to the unexpressed viral capsid protein and eventually not support effective clinical response against diseases (Chabeda et al., 2018; Dadar et al., 2018; Frazer, 2004; Munger et al., 2004).

Alternatively, other strategies are also utilized to target HR-HPVs and LR-HPVs (Dadar et al., 2018; Jung et al., 2015; Kennedy et al., 2014). Like, some studies show the applications of small interfering RNAs (siRNAs) in silencing HPV E6/E7 oncogenes and to kill HPV positive cancer cells (Chang et al., 2010; Jung et al., 2015). Concurrently, clustered regularly interspaced short palindromic repeats (CRISPR)/ CRISPR-associated proteins (Cas) approach utilizing single guide RNAs (sgRNAs) can also be successfully applied to inactivate viral oncogenes and inhibit tumor progression (Hu et al., 2014; Kennedy et al., 2014; Zhen and Li, 2017).

Further, some computational resources dedicated to papillomaviruses were also developed in past few years. One of the most comprehensive resources for papillomavirus studies is the Papillomavirus Episteme (PaVE): a papillomavirus genome database. This database comprises genomes with visualization and analysis tools. It delivers a complete catalog and annotation of papillomavirus genomes. Additionally, it also provides variant, protein structures, transcript, and taxonomy information (Van Doorslaer et al., 2017). Likewise, a knowledgebase "HPVbase" for the three major HPV mediated events, i.e., integration events, HPVs methylation

patterns and miRNAs aberrant expression as a potential biomarkers for HPV associated carcinomas was developed (Gupta and Kumar, 2015, 2016; Kumar Gupta and Kumar, 2015). Another resource human papillomavirus T cell Antigen Database (HPVdb) was developed that provide list of antigens and verified T-cell epitopes (Zhang et al., 2014). To the best of our knowledge, a resource devoted to HPVs therapeutics and epitome is lacking. Thus, we have developed an integrated web-based resource, *HPVomics* (http://bioinfo.imtech.res.in/manojk/hpvomics/) to better understand and provide different putative therapeutic candidates and solutions (Gupta and Kumar, 2020). We hope that this resource will be useful to hasten the anti-HPV research.

Materials and Method

Genomic data collection and curation

There are 210 HPV types were known (http://www.hpvcenter.se), among which 182 types are having the complete genome sequences available that were categorized into five distinct genera namely *Alpha*, *Beta*, *Gamma*, *Mu*, and *Nu* (de Villiers, 2013). In this study, we have predominantly focused on all demarcated HR-HPVs (16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 73, 82) and LR-HPVs (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81). The complete genomic information of distinct HPV types can be retrieved, searched and filtered from the genomic section on the web resource. To gain the correct annotation and updated genomic information Papillomavirus Episteme (PaVE) resource (Van Doorslaer et al., 2017) was utilized along with NCBI and International Human Papillomavirus Reference Center (http://www.hpvcenter.se/html/refclones.html). A total of 18 HR-HPVs and 11 LR-HPVs were included and utilized for the therapeutically oriented analyses (**Table 12**). We have analyzed each gene (i.e. E6, E7, E1, E2, E4, E1^E4, E8^E2, E5, L2, L1) sequences at nucleotide (nt) and amino acid (aa) level. Along with experimental and putative therapeutic regimens, codon usage bias among all the HPVs was also analyzed. HPV genomes were analyzed using the codon usage program from the sequence manipulation suite (Stothard, 2000). Additionally, we have also analyzed codon distribution (rare and preferred) among different HPVs using the Anaconda program (Moura et al., 2005).

Table 12. List of high-risk and low-risk HPVs utilized for the exploration of putative therapeutic and vaccine regimens

Sr. no.	Accession no.	HPV type	Species	Length (bp)	Year
HR-HPVs					
1	NC_001526.4	16	Alpha-9	7906	1984
2	NC_001357.1	18	Alpha-7	7857	1984
3	NC_001583.1	26	Alpha-5	7855	1985
4	J04353.1	31	Alpha-9	7912	1985
5	M12732.1	33	Alpha-9	7909	1985
6	X74477.1	35	Alpha-9	7879	1986
7	M62849.1	39	Alpha-7	7833	1987
8	X74479.1	45	Alpha-7	7858	1986
9	M62877.1	51	Alpha-5	7808	1987
10	X74481.1	52	Alpha-9	7942	1987
11	X74482.1	53	Alpha-6	7859	1987
12	X74483.1	56	Alpha-6	7845	1987
13	D90400.1	58	Alpha-9	7824	1988
14	X77858.1	59	Alpha-7	7896	1989
15	U31794.1	66	Alpha-6	7824	1981
16	X67161.1	68	Alpha-7	7822	1981
17	X94165.1	73	Alpha-11	7700	1993
18	AB027021.1	82	Alpha-5	7870	1997
LR-HPVs					
1	NC_001355.1	6	Alpha-10	7996	1984
2	M14119.1	11	Alpha-10	7931	1984
3	X74478.1	40	Alpha-8	7909	1987
4	M73236.1	42	Alpha-1	7917	1987
5	AJ620205.1	43	Alpha-8	7975	1987
6	U31788.1	44	Alpha-10	7833	1987
7	NC_001676.1	54	Alpha-13	7759	1987
8	U31793.1	61	Alpha-3	7989	1989
9	U21941.1	70	Alpha-7	7905	1993
10	X94164.1	72	Alpha-3	7989	1993
11	AJ620209.1	81	Alpha-3	8070	1996

HPVs putative therapeutic solutions

Vaccine epitopes (HPV Epitome)

All HR- and LR-HPVs proteins were utilized for the purpose of identifying peptides that may induce immune response against the HPVs; 9-mer or 10-mer peptides of HPV proteins were identified. Furthermore, these peptides were analyzed for their immune potential as also mentioned and applied previously (Gupta et al., 2016). Distinctive class of epitopes, i.e., T-cell epitopes (major histocompatibility complex (MHC) class I and MHC class II binders), B-cell epitopes and CTL epitopes were analyzed.

MHC class I and MHC class II binding predictions were performed using the IEDB Analysis Resource Consensus tools (Kim et al., 2012). For the prediction of MHC-I binders IEDB recommended consensus method (ANN, SMM and CombLIb) was employed. Based on the IEDB guideline, IC values ≤50nM are considered as high affinity, here we have used the more stringent selection criteria of IC50 ≤40nM. In addition to this (IC50 values), percentile rank (small numbered percentile rank indicates high affinity) is also provided. Likewise, for the prediction of MHC-II binders IEDB recommended consensus method (NN-align, SMM-align, CombLib and Sturniolo) was employed. In the consensus approach, combination of any three out of the four methods were utilized if suitable model is available otherwise NetMHCIIpan is used (Kim et al., 2012). Percentile rank (small numbered percentile rank indicates high affinity) for each peptide is also provided and used for the ranking.

Furthermore, B-cell 9-mer epitopes (linear) were also predicted using LBtope (Singh et al., 2013) methods. To identify the favourably potent and reliable epitopes, a cut-off of 70% was selected for the prediction. Further, to gain more confidence only confirms dataset model was utilized, which is developed using the epitopes that are verified at least by two studies. Similarly, putative CTL epitopes were also detected via CTLPred (Bhasin and Raghava, 2004) algorithm built on artificial neural network (ANN) and support vector machine (SVM) modules. For prediction of CTL epitopes, consensus (ANN+SVM) prediction approach was utilized and top three epitopes for each protein of HR-HPVs and LR-HPVs were catalogued. The comprehensive compendium of potentially useful all categories of epitopes were compiled and provided at web resource.

In addition to this, mining and collection of experimentally verified epitope data was also performed utilizing Immune Epitope Database (IEDB) (Kim et al., 2012; Vita et al., 2015) and systematic curation and integration of different data was performed. Additionally, an interactive epitope map for the HPV16, 18, 33 and HPV11 proteins was also constructed.

Anti-viral peptides (AVPs)

We have also extracted and adapted HPV specific experimentally verified antiviral peptides from AVPdb (Qureshi et al., 2014d) resource and provided with the peptide properties.

Small interfering RNAs (siRNAs)

Furthermore, all the genes of HR- and LR-HPVs were targeted to design and identify potentially effective siRNAs. We have utilized the VIRsiRNApred (Qureshi et al., 2013b) software for the prediction of effective siRNAs against these HPVs with the efficacy cut off of ≥50% along with the off-targets information. To provide easy access to this, protein wise representation was adopted and provided on resource.

Single guide RNAs (sgRNAs) identification

Likewise, gene-wise sequences of HR and LR-HPVs were also screened and evaluated for the potent sgRNAs utilizing an integrated pipeline "geCRISPR" for the identification of possible sgRNAs on the basis of protospacer adjacent motif (PAM) (Kaur et al., 2016). On both strands (forward and reverse) of genomic sequence "NGG" motifs were scanned and then highly potent putative sgRNAs or CRISPR targets were identified and provided.

HPVepi: HPV epitome prediction algorithm

Experimentally verified non-redundant epitopes were retrieved for each class of epitopes from IEDB v3.0 dated 21.12.2018. In total, 1001 peptides of HPV B-cell of which 491 are epitopes (positive^{P}) and 510 are non-epitopes (negative^{N}); 507 T-cell MHC-I peptides of which 268^{P} are epitopes and 239^{N} are non-epitopes; 665 T-cell MHC-II peptides of which 326^{P} are epitopes and 339^{N} are non-epitopes were obtained (**Table 13**). Further, for the algorithm development, peptides of length between 5-40 amino acids (aa) were considered. Peptides which are too short i.e. less than 5 aa and too large i.e. more than 40 aa length were removed to form a working dataset from

each epitope category. Overall, after data curation based on length filtering, B-cell positive dataset 470^{P}, T-cell MHC-I negative dataset 237^{N}, T-cell MHC-II positive dataset 318^{P} and negative dataset 332^{N} was remained (**Table 13**).

Overall, for B-cell 980 peptides (470^{P}+510^{N}), T-cell MHC-I 505 peptides (268^{P}+237^{N}), and T-cell MHC-II 650 peptides (318^{P}+332^{N}) were used for the algorithm development. For B-cell prediction algorithm development from 980 sequences, we have randomly extracted 196 sequences (20%) as independent/validation datasets (V196); other remaining 784 sequences were used for the 5nCV training/testing datasets (T/T-784). For T-cell MHC-I prediction algorithm, we randomly extracted 101 sequences (20%) as independent/validation datasets (V101); while the remaining 404 peptide sequences were utilized as training/testing datasets (T/T-404). Likewise, For T-cell MHC-II prediction algorithm, we randomly extracted 130 peptides (20%) as independent/validation datasets (V130); while the remaining 520 peptide sequences were utilized as training/testing datasets (T/T-520) (**Table 13**).

To train and test on the different peptide sequence features, SVM^{light} (v6.02) software package (http://svmlight.joachims.org) was utilized as also described in various studies (Dar et al., 2016; Kaur et al., 2016; Rajput et al., 2015; Thakur et al., 2012a). Different sequence features i.e. amino acid composition (AAC), di-peptide composition (DPC), binary 5N5C profile (BIN5N5C) (positional profile), and hybrids of them were used for the model development and 5nCV. The frequency of each amino acid in the peptide sequence makes the compositional profile of sequences. It also useful in order to make fix length vector irrespective of variable length sequences necessary for the machine learning techniques (MLTs). We have utilized the AAC and DPC profiles which forms the vector of length 20 and 400, respectively. We also studied profile of positions of different amino acids (binary profile) from peptide sequences. As the peptides are of variable lengths, we have considered 5 aa from each end i.e. N and C terminal (as the smallest peptide is of length 5 aa) to make the vector of fixed length 10. In binary profile, occurrence of each amino acid at 5N and 5C positions was studied. Additionally, for the model development, individual features i.e. AAC, DPC, and BIN5N5C were also used in combination as hybrid approach to form the different hybrid features and models.

Table 13. Overview of data utilized in HPVepi algorithm

	Epitope Classes		
	B-cell	**T-cell (MHC-I)**	**T-cell (MHC-II)**
All	1001/980*	507/505*	665/650*
+ve	491/470*	268	326/318*
-ve	510	239/237*	339/332*
Training/Testing (T/T) dataset	784	404	520
Independent validation (V) dataset	196	101	130

*Curated peptide data after length-based filtering utilized in the algorithm

Further, based on the optimal and maximal result using different parameters, final deployable predictive model and classifier is developed. Performances of models were evaluated based on the sensitivity, specificity, accuracy, Mathew's correlation coefficient (MCC), receiver operating curve (ROC) etc. Independent validation of the developed final classifier is also performed.

Sensitivity refers and represents the model ability to correctly predict the positive epitopes from the actual positive epitopes.

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

Where, TP is true positive (correctly predicted positive epitopes); FN is false negative (falsely predicted negative epitopes).

Specificity measures the test ability to rightly predict negative epitopes from actual negative epitopes.

$$Specificity = \frac{TN}{TN + FP} \times 100$$

Where, TN is true negative (correctly predicted negative epitopes); FP is false positive (falsely predicted positive epitopes).

Accuracy represents the percentage of correctly predicted epitopes from the complete data (positive and negative).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

Mathew's correlation coefficient (MCC) refers to the correlation between the observed and the predicted classification. MCC value ranges from -1 to +1 where, -1 represent negative correlation, 0 shows random correlation and +1 signify perfect correlation.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100$$

Receiver operating characteristic (ROC) is a threshold independent criterion to study the performance of predictive models. ROC curve can be plotted between the true positive rate (TPR) i.e. sensitivity against the false positive rate (FPR), i.e., 1-specificity. The complete workflow of HPVepi algorithm is represented in **Figure 41**.

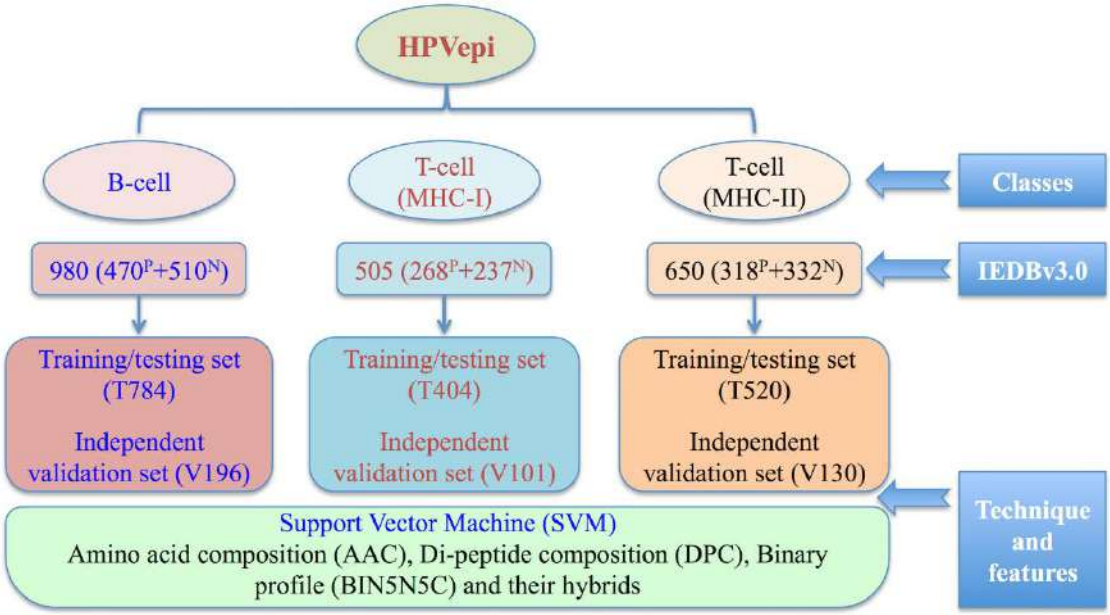


Figure 41. Computational workflow of HPVepi algorithm.

Resource and web server implementation

All the genomic data files were converted using a Perl in-house script in gene feature files (GFF3 format) and descriptive sections related to genomic annotation or regulatory information configured using Jbrowse (Skinner et al., 2009), a java-script based browser to visualize interactively as also implemented in previous studies (Gupta et al., 2016; Khan et al., 2016). The web interface and back-end are constructed using the combination of different web programming and scripting languages i.e. Perl, PHP, Java-scripts, and HTML etc. HPVepi algorithm is also integrated on the resource developed using SVM^{light} package. The complete system is

hosted using the open source LAMP server to utilize Apache, MySQL, and PHP on Linux environment.

Results and discussion

HPVomics overview

HPVomics (http://bioinfo.imtech.res.in/manojk/hpvomics/) is a web-based HPV therapeutic and genomic resource; comprehensively provide and especially dedicated towards putative therapeutically important solutions along with genomic information. It is classified in different sections represent individual components mainly therapeutics (i.e. siRNAs, sgRNAs, antiviral peptides etc.), vaccine epitopes (IEDB epitopes, MHC-I and -II binders, B-cell, cytotoxic T lymphocytes (CTL) epitopes), genomes, genome browser, epitope map, HPVepi prediction algorithm and tools (**Figure 42**). It has well-designed and an easy user interface for the interactive visualization and evaluation.

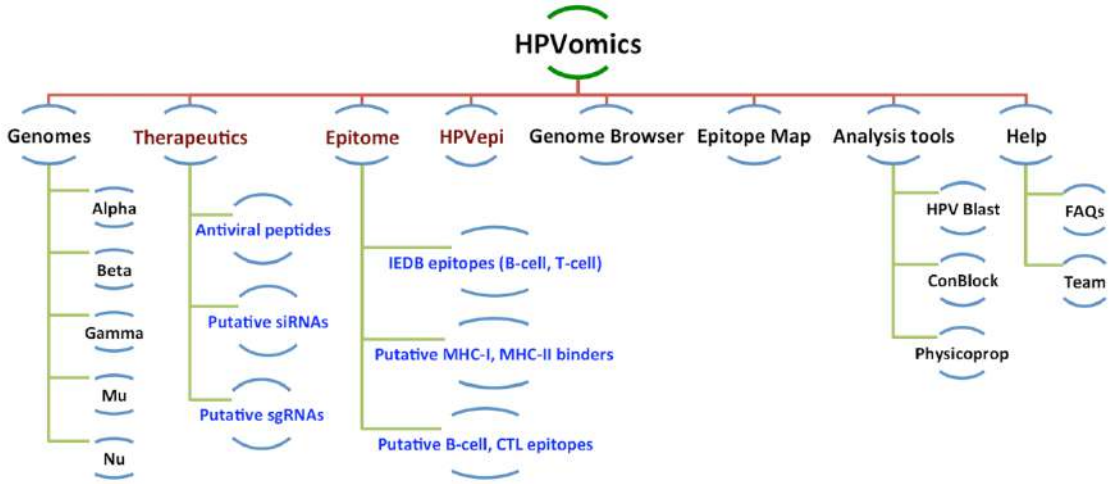


Figure 42. HPVomics architecture

Potential therapeutic solutions

HPVs vaccine candidates

In this study, efforts were made to elucidate promising HPV T-cell (MHC-I and MHC-II binders), B-cell and CTL epitopes. Overall, all the protein targets (E6, E7, E1, E2, E4, E1^E4, E8^E2, E5, L1, L2) were utilized from the HR-HPVs (18) and LR-HPVs (11) (**Table 12**) to identify the putative vaccine epitome. Additionally, experimentally verified B-cell and T-cell epitopes were also retrieved from IEDB and reported.

In total, 604 potential B-cell epitopes from HR-HPVs and 330 from LR-HPVs, with the meta-information i.e., their sequence, predicted score and B-cell confidence percent value were specified. These epitopes exhibiting highest potency and confidence percentage are proposed as potential targets. Likewise, putative 467 and 304 CTL epitopes from HR-HPVs and LR-HPVs respectively, were catalogued and integrated in the resource. The comprehensive knowledge such as epitope sequence, start coordinate, end coordinate and allele information are provided. Number of B-cell and CTL epitopes specific to each HR- and LR-HPVs was specified in **Figures 43 and 44,** respectively. Furthermore, the potential high affinity repertoire of MHC class I and II binders are also provided for each protein from both HR- and LR-HPVs. Overall, 4228 MHC class I and 3712 MHC class II focusing epitopes pertaining to eighteen HR-HPVs were reported. Likewise, 2498 MHC-I and 2512 MHC-II epitopes from eleven LR-HPVs were classified and provided. Number of protein-wise promiscuous MHC binders is depicted in **Figure 45.** Overall, this prospective and recommended vaccine epitome could be useful and assist in effective vaccine design.

Further, we have also retrieved and reported the experimentally verified B-cell and T-cell epitopes using IEDB. Overall, 1687 B-cell epitope entries were identified that primarily belong to HPV16 (1282). Likewise, 1823 T-cell epitope entries were catalogued, which are mainly pertain to HPV16 (1328), HPV18 (167) and HPV11 (113). It provides complete information regarding the epitopes like HPV type, gene region, epitope sequence, epitope type, start, end, length, method, assay group, cell type, alleles, MHC class, reference etc. HPV type wise numbers of experimentally proven T-cell and B-cell epitope entries were represented in line diagram (**Figure 46A-B**). Additionally, we have also represented interactive browser of epitopes. In Epitope map (**Figure 47**), data is classified according the viral proteins with priority given to the HPV oncoproteins (E6 and E7). It contains information such as IEDB id, start and end coordinates with gene region, epitope length, epitope sequence, and HLA allele type (**Figure 47**).

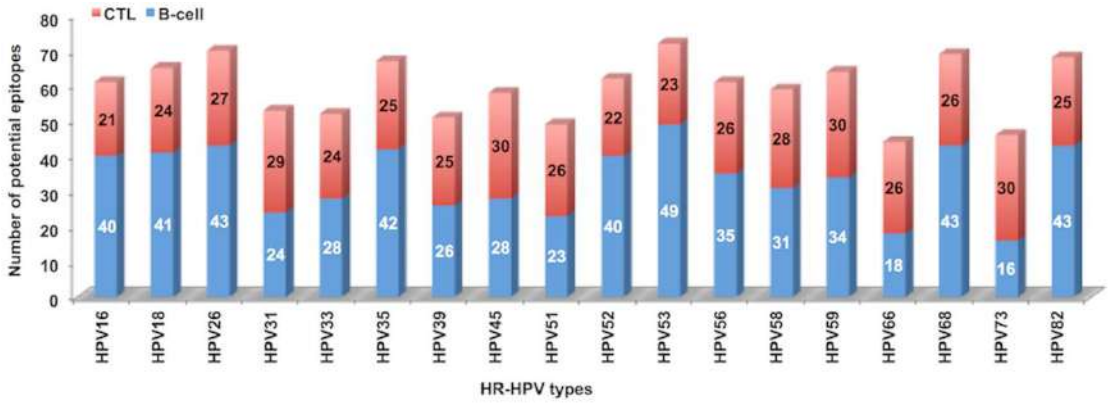


Figure 43. Number of B-cell and CTL epitopes from HR-HPVs

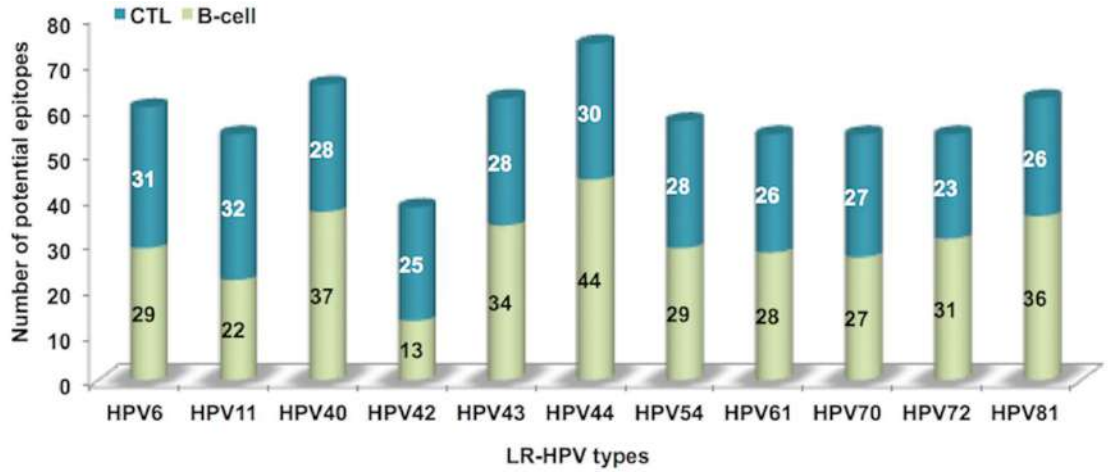


Figure 44. Number of B-cell and CTL epitopes from LR-HPVs

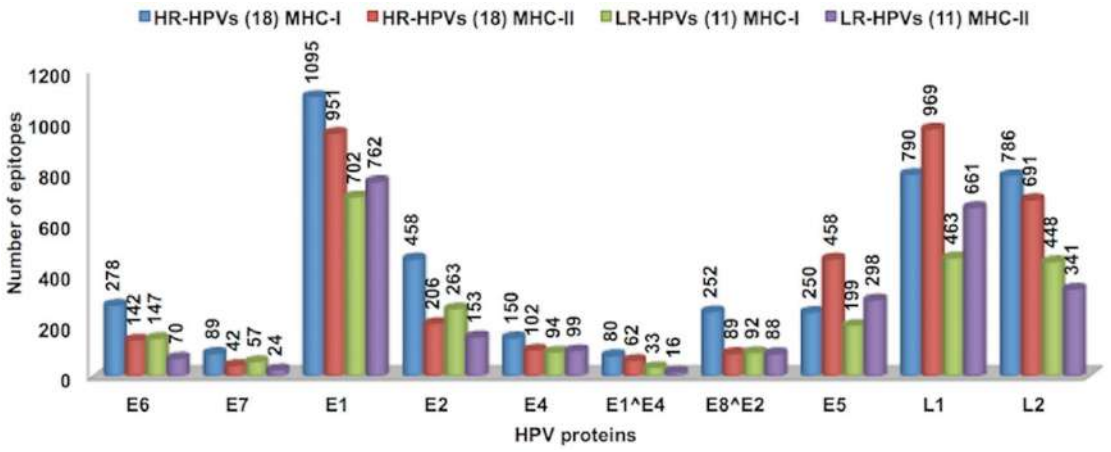


Figure 45. Number of MHC-I and MHC-II binding epitopes from HR- and LR-HPVs

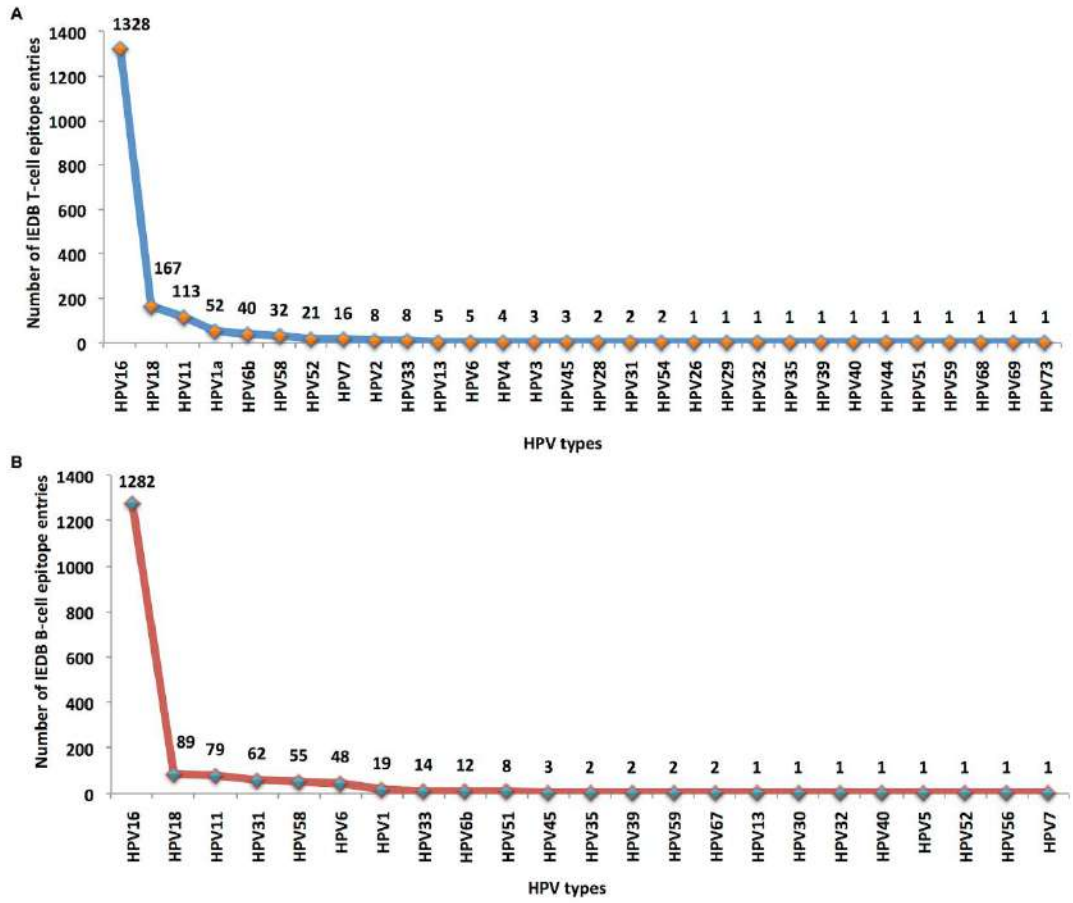


Figure 46. Number of experimentally proven IEDB epitopes. (A) T-cell and (B) B-cell

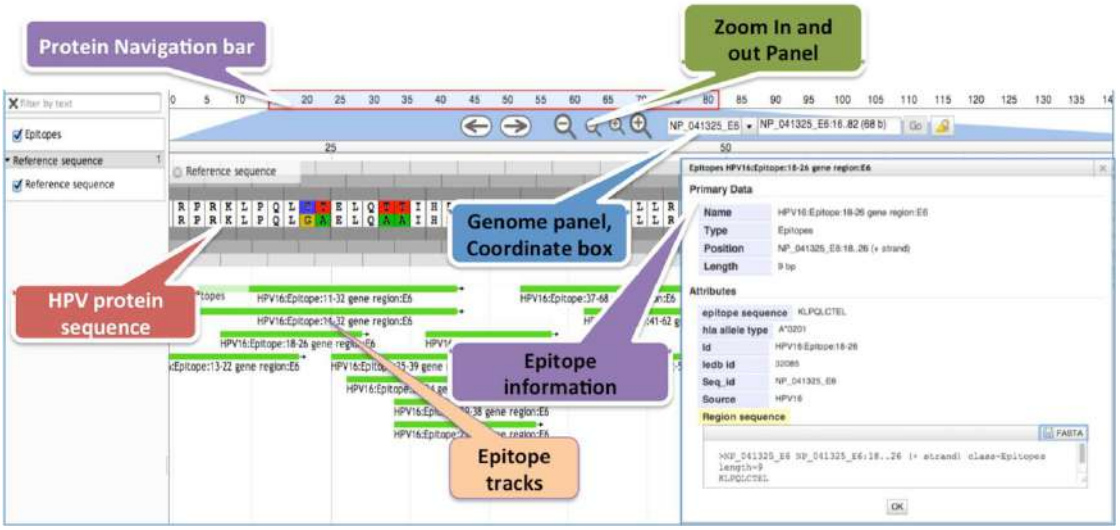


Figure 47. Screenshot depicting Epitope map tracks from HPVomics. Epitopes (showing in green) are mapped on the reference protein sequence E6 (NP_041325). User can enlarge reference track to visualize sequence and move from upper coordinate scale. Epitope information (in inset) can be visualized by selecting epitope track

Anti-viral peptides

Importantly, anti-viral peptides could be very effective in the effort to combat HPV infection. In this study, we have also compiled experimentally proven natural and synthetic anti-viral peptides for the HPVs. Overall, 24 natural AVPs and 15 synthetic AVPs were recorded. The complete information such as peptide sequence, length, source, inhibition/IC_{50} value, cell line information, Uniprot ID, target, assay, and cross references were provided.

RNA based therapeutics

Small interfering RNAs (siRNAs)

RNA based therapeutic regimens can also be used for the effective targeting of HPVs infection especially for oncogenes. Here, the aim is to provide complete spectrum of potentially effective siRNAs. For the same, we have catalogued 1567 effective siRNAs (≥50% efficacy) from eighteen HR-HPVs and 1162 siRNAs pertaining to eleven LR-HPVs for the effective inhibition of target mRNAs. We have utilized the VIRsiRNApred algorithm to predict the efficacy percentage. These can also be used to invoke immune system (immunomodulatory siRNAs). Web server delivers exhaustive information including siRNAs sense and antisense sequence, start and end position on genome, HPV target region, inhibition percentage and seed based off-targets. **Figure 48** illustrating the number of siRNAs belongs to different HPV genes from both HR- and LR-HPVs. Overall, 195 potential siRNAs targeting different genomic regions of HR-HPV 16 showed 50 percent or more silencing efficacy shown in circular diagram (**Figure 49**) powered by Circos software (Krzywinski et al., 2009).

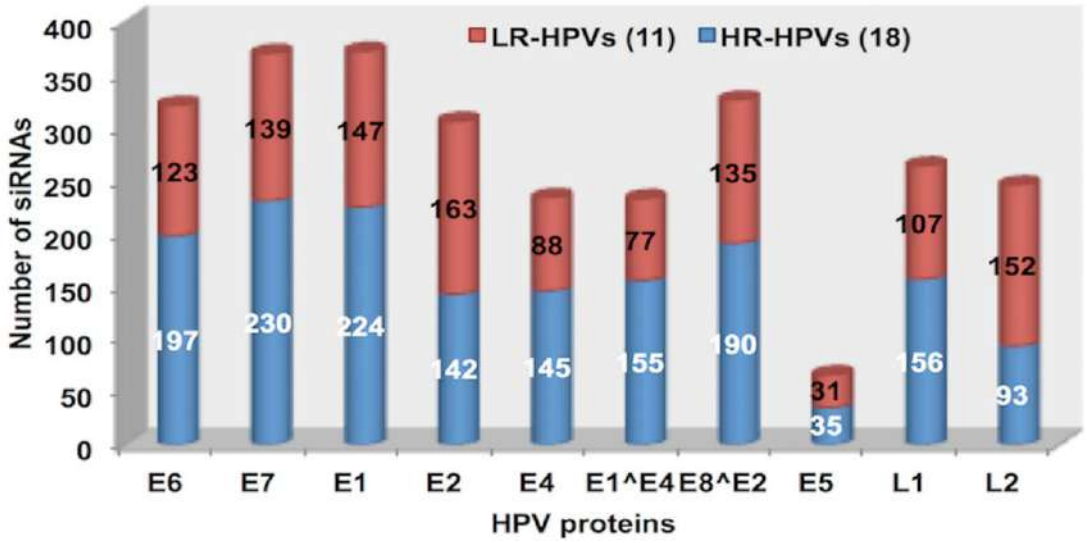


Figure 48. Number of potentially effective siRNAs against HPV genes associated to HR-HPVs and LR-HPVs

Single guide RNAs (sgRNAs)

Likewise, recently it is also shown that CSISPR/Cas technology can have great applications in effective targeting of genomes. From analysis, we have represented 1451 and 1094 effective sgRNAs (≥40% efficacy) from eighteen HR-HPVs and eleven LR-HPVs, respectively. Tabular illustration describes HPV type, target region on HPV genome, sgRNA sequences, PAM motif, strand (sense/antisense), start and end of the 23-residue sgRNAs, G+C content and predicted genome editing efficiency percentage. User can search sgRNAs using distinct criteria and genomic regions. Highly potent sgRNAs (efficiency ≥50%) of HR-HPVs, i.e., 16 and 18 were depicted in Circos diagram (**Figure 50**). Distribution of sgRNAs from the distinct HPV genes from HR- and LR-HPVs is shown in **Figure 51**.

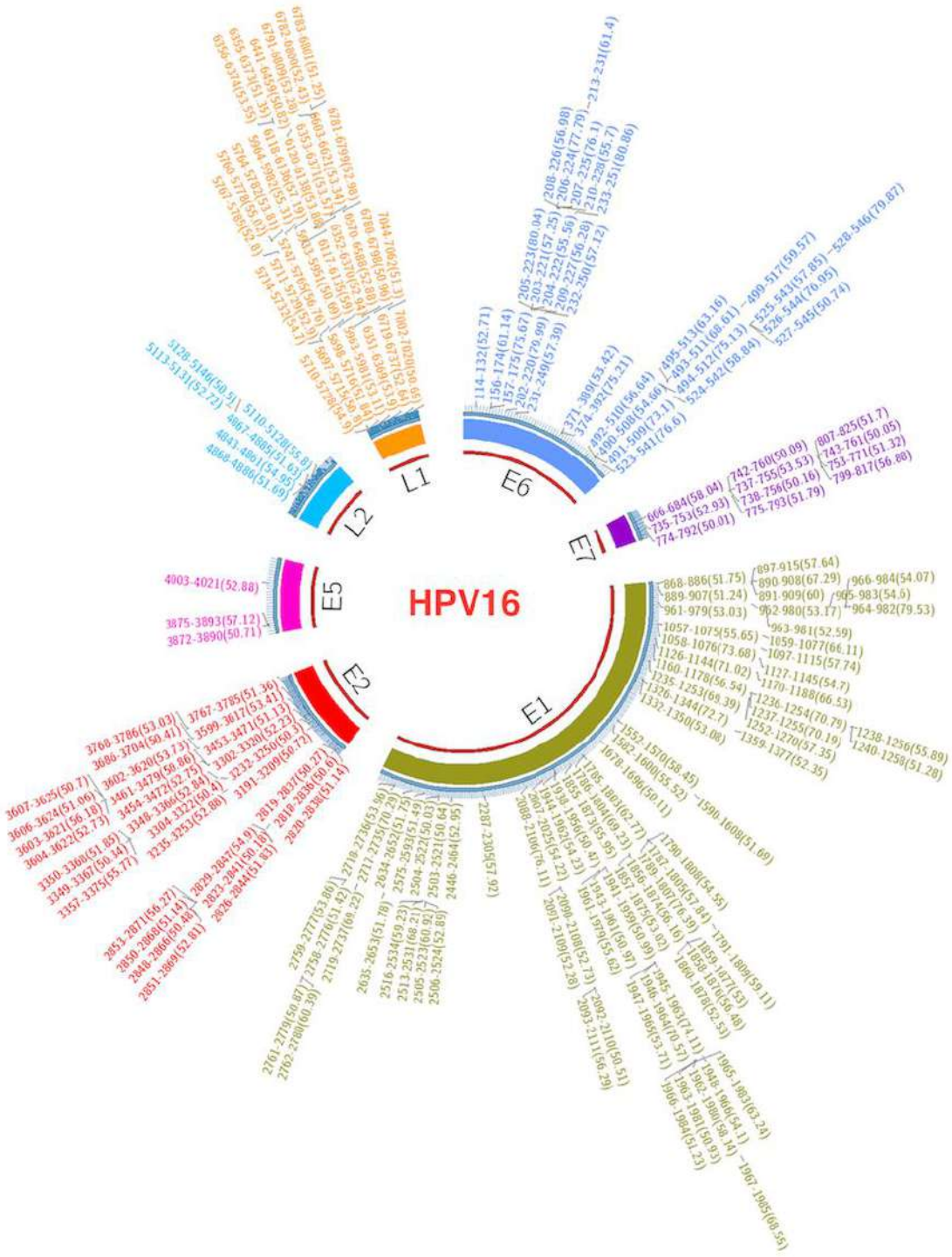


Figure 49. Circular plot representing putative efficient siRNAs (Efficacy ≥50%). HPV16 gene wise start and end of siRNAs with its efficacy were shown in plot

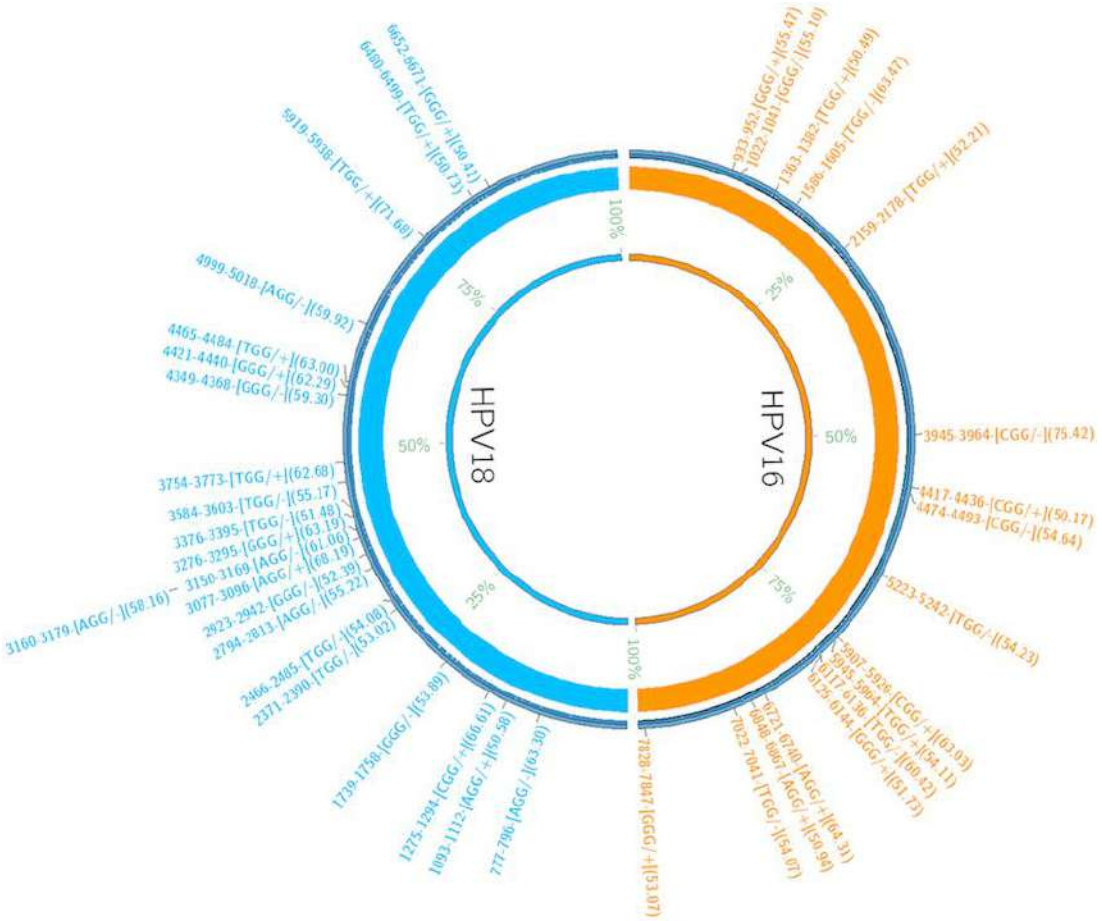


Figure 50. Circos plot depicting putative efficient sgRNAs (Efficacy ≥50%) of HPV 16 and 18

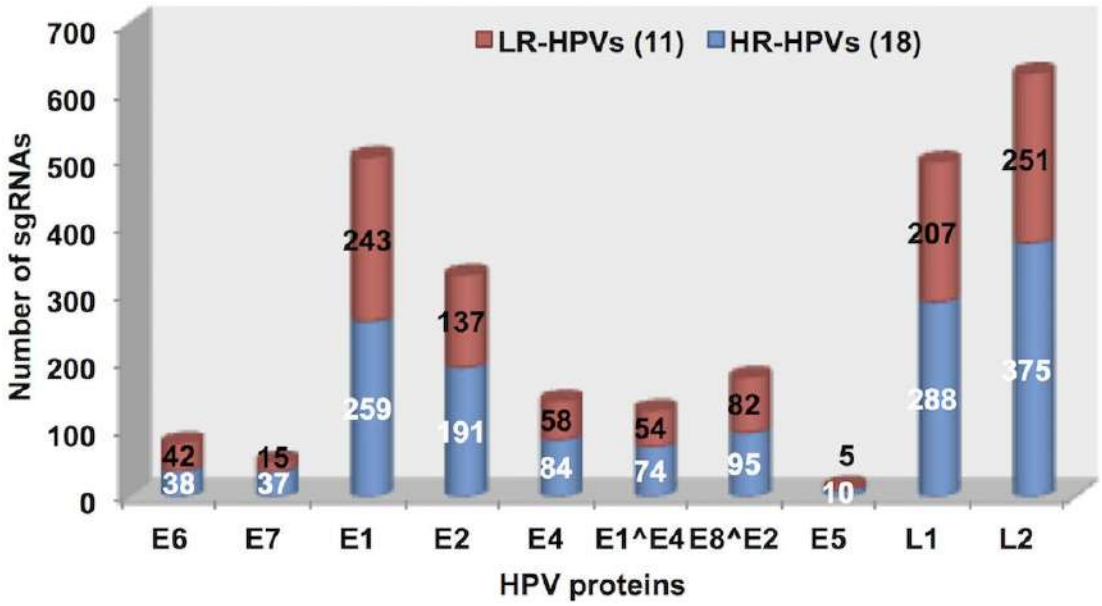


Figure 51. Number of potentially effective sgRNAs against distinct HPV genes from HR- and LR-HPVs

HPV genomes, browser and codon usage

All genomic information of HPVs was compiled and content were provided in tabular manner with searching and filtering facility. This provides different information such as HPV type, genus, species, Genbank accession, length, fasta and genbank files, refseq accession, submission year, literature reference, codon usage analysis etc. Additionally, to navigate throughout the HPV genomes, we constructed an "HPVomics genome browser" for interactive annotation visualization. Various tracks include HPV reference genome, Genes, CDS, mRNA, exon, CpG islands, promoter, TATA box, CAAT signal, 5' UTR, repeat region, protein binding site, polyA signal sequence, and secondary structure information (**Figure 52**). Further, we have also analyzed rare (blue color) and preferred codons (black color) represented in the form of histograms for each genome. Complete catalogue of codon distribution is also accessible at HPVomics resource.

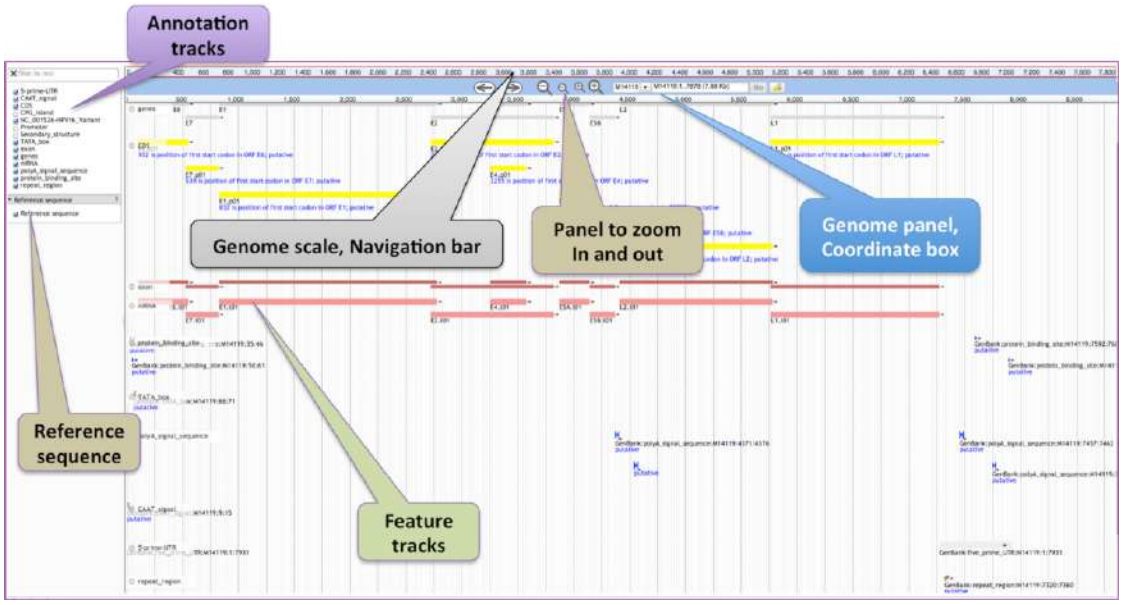


Figure 52. Overview of HPVomics genome annotation browser. The upper panel shows the positional scale (ruler) to navigate through genomes along with HPV reference sequence. Distinct annotation features were shown in separate color blocks. Semantic navigation and zooming provide interactivity to browser.

HPVepi: HPV epitome prediction algorithm

We have developed a computational algorithm, HPVepi, for the prediction of HPV B-cell, T-cell (MHC-I) and T-cell (MHC-II) epitopes. Prediction algorithm for all three arms of immunity is developed utilizing experimentally verified non-redundant peptides (epitopes and non-epitopes) from the IEDB v3.0. We have generated different predictive models and performance was evaluated using several peptide sequence features employing support vector machine (SVM).

Performance of HPV B-cell prediction method during 5-fold cross validation (5nCV) as well on independent validation

In HPVepi algorithm for each class of prediction method different sequence features i.e. amino acid composition (AAC), di-peptide composition (DPC), binary 5N5C profile (BIN5N5C) (positional information), and hybrids of them were used in both 5nCV and independent validation. These sequence features were used on all three random training/testing dataset (T784) and performance was evaluated. Further, performance of models was also measured on the independent validation set (V196).

On composition profile i.e. amino acid composition (AAC), di-peptide composition (DPC) we have achieved accuracy and Mathew's correlation coefficient (MCC) of 71.56%, 74.23% and 0.43, 0.48 correspondingly. Further, performance of position based binary feature BIN5N5C also evaluated that attained accuracy of 66.96% and 0.34 MCC that do not perform better than the individual compositional features. Furthermore, we have also developed hybrid predictive models i.e. AAC+DPC and AAC+DPC+BIN5N5C. We have achieved a maximum accuracy of 74.74% and correlation of 0.49 in both the hybrid models. Overall, in B-cell prediction method both the hybrids performed equally well and better than the individual features. We have deployed AAC+DPC+BIN5N5C hybrid model for the web server implementation. Detailed results on all random sets on different features are shown in **Table 14**. Apart from 5-fold cross validation; we have also performed the evaluation of predictive models on the independent validation dataset (V196). Overall, all the models on different features performed equally well on validation set. The best performing predictive method i.e. AAC+DPC+BIN5N5C hybrid model shows equivalent performance with accuracy and MCC of 73.47%, 0.47 respectively (**Table 14)**.

Performance of HPV T-cell (MHC-I) prediction method during 5-fold cross validation (5nCV) and on independent validation

We have utilized different features as also mentioned above to determine the efficiency of predictive models for the prediction of T-cell (MHC-I) epitopes. Best possible predictive models are chosen considering both 5nCV as well as independent validation values. We have achieved accuracy of 58.91%, 56.93% and MCC of 0.18, 0.14 on compositional features i.e. ACC and DPC respectively during 5nCV on the training/testing set (T404) and in independent validation (V101) 57.43% accuracy, 0.15 MCC on AAC and 56.44% accuracy and 0.13 MCC on DPC. We obtained accuracy of 57.92% and MCC of 0.16 on BIN5N5C model during 5nCV and during validation 58.42% accuracy and 0.16 MCC. Further, during 5nCV, we have achieved accuracy and MCC of 57.92%, 60.64% and 0.16, 0.21 respectively for AAC+DPC and AAC+DPC+BIN5N5C hybrid models. During independent validation, on AAC+DPC we have obtained 61.39% accuracy and 0.23 MCC. Likewise, on AAC+DPC+BIN5N5C hybrid model, we have achieved 57.43% accuracy and MCC of 0.15. Detailed result on different features is presented in **Table 15**. All the models performed fairly well on independent dataset (**Table 15)**. Amongst all, the best doing predictive AAC+DPC hybrid model during independent data is implemented on web server.

Performance evaluation of HPV T-cell (MHC-II) prediction method during 5-fold cross validation (5nCV) and on independent validation

We have also developed method to predict the T-cell MHC-II class epitopes for HPVs. In 5nCV on training/testing set (520), on compositional profile i.e. AAC and DPC we have obtained accuracy and MCC of 61.15%, 61.92% and 0.22, 0.24 correspondingly. Likewise, performance of BIN5N5C model is also calculated and provided the accuracy of 55.38 and 0.11 MCC, which is underperforming than the distinct compositional features. Additionally, we have further evaluated hybrid models i.e. AAC+DPC, AAC+DPC+BIN5N5C. We have achieved 61.35% accuracy and MCC of 0.23 in AAC+DPC hybrid model. Likewise, on AAC+DPC+BIN5N5C hybrid model we obtained accuracy of 61.15% and 0.22 MCC. We have deployed best performing DPC model for the web server implementation. Complete result is shown in **Table 16**. Further, on independent validation set (V130), all the models performed equivalent. On the top performing predictive DPC feature, we have achieved the accuracy of 68.46% and 0.37 MCC (**Table 16**).

Table 14. Performance of predictive models during 5-fold cross validation (5nCV) on training/testing data (T784) and on independent validation set (V196) for HPV B-cell prediction algorithm on different features

Features	T784						V196					
	Th	SEN	SPEC	ACC	MCC	AUC	Th	SEN	SPEC	ACC	MCC	AUC
AAC	0	70.74	72.06	71.43	0.43	0.761	-0.1	68.09	71.57	69.90	0.40	0.753
DPC	0	73.67	74.75	74.23	0.48	0.796	-0.1	71.28	70.59	70.92	0.42	0.779
AAC+DPC	-0.1	74.73	74.75	74.74	0.49	0.796	-0.1	70.21	72.55	71.43	0.43	0.785
BIN5N5C	-0.1	67.82	66.18	66.96	0.34	0.715	0	65.96	63.73	64.80	0.30	0.729
AAC+DPC +BIN5N5C	**-0.1**	**74.47**	**75**	**74.74**	**0.49**	**0.796**	**0**	**74.47**	**72.55**	**73.47**	**0.47**	**0.777**

AAC, amino acid composition; DPC, di-peptide composition; BIN, binary; Thres, threshold; SEN, sensitivity; SPEC, specificity; ACC, accuracy; MCC, Matthew's correlation coefficient; AUC, area under curve

Table 15. Performance of predictive model during 5-fold cross validation (5nCV) on training/testing data (T404) and on independent validation set (V101) for HPV T-cell (MHC-I) prediction algorithm on different features

Features	T404						V101					
	Th	SEN	SPEC	ACC	MCC	AUC	Th	SEN	SPEC	ACC	MCC	AUC
AAC	0.4	59.81	57.89	58.91	0.18	0.608	0.4	57.41	57.45	57.43	0.15	0.601
DPC	0	57.94	55.79	56.93	0.14	0.577	0.1	55.56	57.45	56.44	0.13	0.622
AAC+DPC	**0**	**59.35**	**56.32**	**57.92**	**0.16**	**0.569**	**0.1**	**59.26**	**63.83**	**61.39**	**0.23**	**0.637**
BIN5N5C	0.1	57.48	58.42	57.92	0.16	0.606	0.1	61.11	55.32	58.42	0.16	0.531
AAC+DPC +BIN5N5C	0.5	57.94	59.47	58.66	0.17	0.602	0.1	61.11	57.45	59.41	0.19	0.635

AAC, amino acid composition; DPC, di-peptide composition; BIN, binary; Thres, threshold; SEN, sensitivity; SPEC, specificity; ACC, accuracy; MCC, Matthew's correlation coefficient; AUC, area under curve

Table 16. Performance of predictive model during 5-fold cross validation (5nCV) on training/testing data (T520) and on independent validation set (V130) for HPV T-cell (MHC-II) prediction method

Features	T520						V130					
	Th	SEN	SPEC	ACC	MCC	AUC	Th	SEN	SPEC	ACC	MCC	AUC
AAC	-0.1	62.6	59.77	61.15	0.22	0.610	0	57.81	59.09	58.46	0.17	0.655
DPC	**-0.2**	**64.57**	**59.4**	**61.92**	**0.24**	**0.624**	**-0.1**	**67.19**	**69.70**	**68.46**	**0.37**	**0.675**
AAC+DPC	-0.1	62.2	60.53	61.35	0.23	0.630	-0.1	64.06	63.64	63.85	0.28	0.660
BIN5N5C	-0.1	54.33	56.39	55.38	0.11	0.552	-0.1	53.12	50.00	51.54	0.03	0.534
AAC+DPC +BIN5N5C	-0.1	61.81	60.53	61.15	0.22	0.630	-0.1	64.06	62.12	63.08	0.26	0.660

AAC, amino acid composition; DPC, di-peptide composition; BIN, binary; Thres, threshold; SEN, sensitivity; SPEC, specificity; ACC, accuracy; MCC, Matthew's correlation coefficient; AUC, area under curve

Comparison of HPVepi with the existing algorithms

Here we also describe the comparison of our algorithm with the existing epitope prediction methods for the prediction of HPV epitome. Up to now there is no specific method developed to predict the HPVs epitopes, and also maximum methods were developed to predict binders and non-binders, however there are few epitope immunogenicity prediction methods are available. These methods are developed specific for individual class of epitopes, i.e., for B-cell, CD4 and CD8 T-cells separately. However, we have developed an integrated web server, HPVepi, which can predict peptide potentiality for all epitome classes i.e. cross capacity to elicit immune response. Further, the existing web servers either prefer or require the fixed length of peptides as input but our algorithm provides flexibility in term of peptide length.

We compared our algorithm separately to different methods for all three categories of epitopes. For the unbiased comparison, we have opted and utilized the independent validation peptide data for each category separately. For B-cell prediction, we have compared our HPV B-cell method with the LBtope (Singh et al., 2013) and BepiPred-2.0 (Jespersen et al., 2017). Further for T-cell MHC I prediction, we compared T cell class I pMHC immunogenicity predictor (Calis et al., 2013) with our method. Likewise for T-cell MHC-II prediction, we compared our method with the CD4 T cell immunogenicity predictor (Dhanda et al., 2018). Our algorithms performed relatively better than the existing servers to predict HPV epitome.

In lbtope, if probability score is greater or equal to 60% (as mentioned on the server), prediction is considered as positive else marked as negative. Contrarily, Bepipred 2.0 provides epitope probability score of individual amino acids in a peptide/protein and based on threshold, it marks each residue above epitope threshold as E. It also requires peptide/protein sequences should be more than or equal to 10 amino acid (aa) in length. To evaluate the performance of bepipred 2.0 on independent data, we have first removed peptides which are less than required length (10 aa). In total, out of 196 peptides, 169 remained. To make the conclusion of positive or negative prediction, we opted a formula i.e. if half of the residue of the peptide is above the threshold (0.5) than given peptide is marked as positive (P) else negative (N).

$$Positive\ (P) = En \geq L/2$$

$$Negative\ (N) = En \leq L/2$$

Where, En is total number of E residues (above threshold) in a given peptide, L is length of a given peptide. Based on the above-mentioned criteria, true positives, false positives, true negatives, false negatives were counted and sensitivity, specificity and accuracy are calculated for both the methods (**Table 17**). Our algorithm shows best performance with 73.47% accuracy on independent validation dataset as compare to both the algorithms i.e. LBtope (61.73%) and Bepipred-2.0 (60.95). However, both the external methods exhibit greater specificity.

Table 17. Performance evaluation of existing B-cell epitope prediction methods on independent validation data

Methods	**Dataset**	**SEN (%)**	**SPEC (%)**	**ACC (%)**
LBtope	196	40.43	81.37	61.73
BepiPred-2.0	169 (>=10 aa)	24.64	86	60.95
HPVepi-B-cell	196	74.47	72.55	73.47

SEN, sensitivity; SPEC, specificity; ACC, accuracy

Further, T cell class I pMHC immunogenicity predictor used to provide prediction score for each peptide. If the prediction score of a peptide is greater than 0, then it is considered as positive prediction else negative. Based on prediction outcome, performance in terms of accuracy is calculated (**Table 18**). Our method provides better accuracy of 61.39% in comparison to 48.51% accuracy of T cell class I pMHC immunogenicity predictor.

Table 18. Performance evaluation of existing T-cell (MHC-I) epitope prediction method on independent validation data

Methods	**Dataset**	**SEN (%)**	**SPEC (%)**	**ACC (%)**
T cell class I pMHC immunogenicity predictor	101	53.7	42.55	48.51
HPVepi_T-cell (MHC-I)	101	59.26	63.83	61.39

SEN, sensitivity; SPEC, specificity; ACC, accuracy

Furthermore, CD4 T cell immunogenicity predictor was evaluated. In this we have used IEDB recommended combined method (7-allele method + immunogenicity method) for the prediction. This also requires sequence to be of length 15-mer or more. We have extracted all the 15-mer peptides from the independent data (V130). Overall, 53 peptides were remained. Server provides two scores i.e. combined score and immunogenicity score. For decision-making, we have utilized the combined score and cut-off of 50% is used to make prediction outcome as positive (≥50%) or negative (<50%). Finally, sensitivity, specificity and accuracy are computed (**Table 19**). Our method shows superior performance with the accuracy of 68.46% on independent data of variable length as compare to CD4 T cell immunogenicity predictor with accuracy of 37.74% on 15-mer peptides (**Table 19**).

Table 19. Performance evaluation of existing T-cell (MHC-II) epitope prediction method on independent validation data

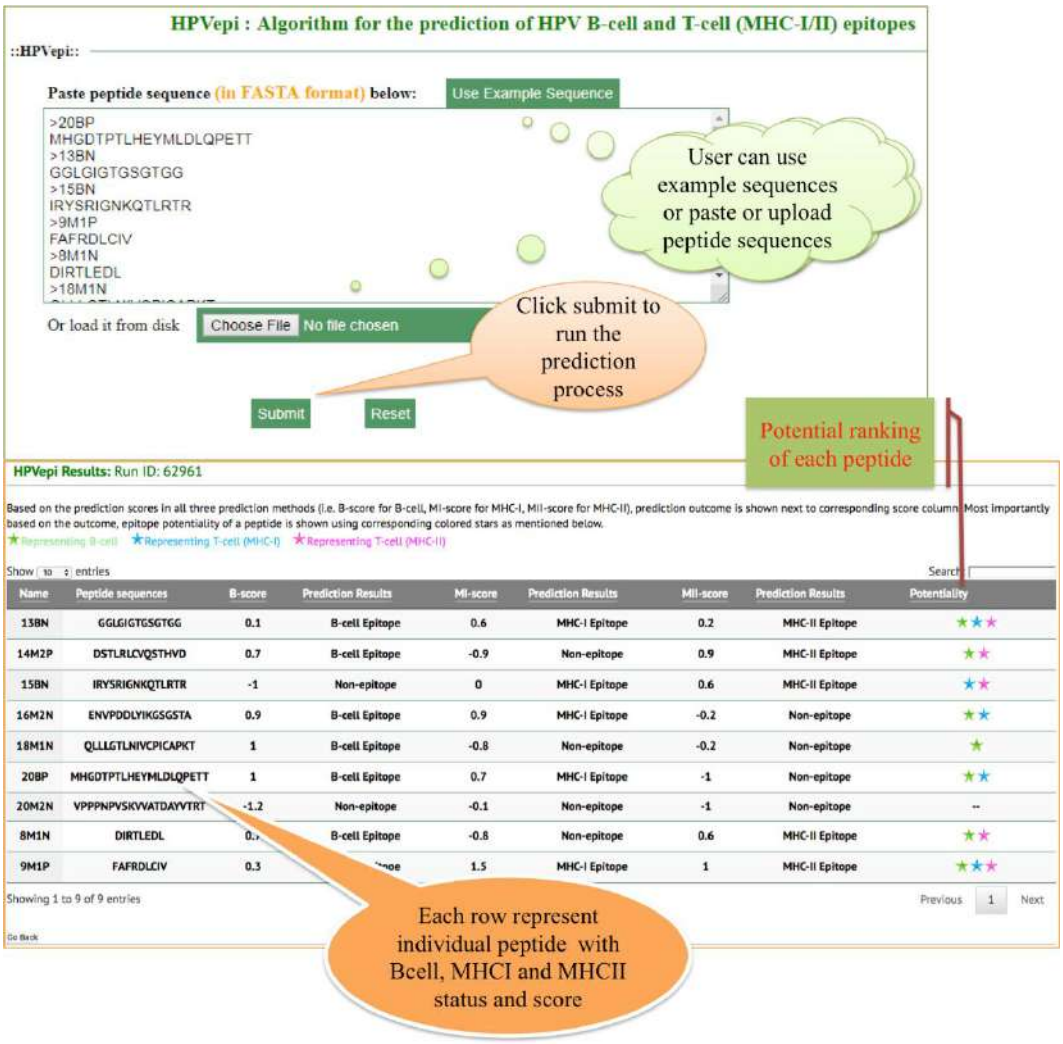
Methods	**Dataset**	**SEN**	**SPEC**	**ACC**
CD4 T cell immunogenicity predictor	53 (Length=15 aa)	50%	21.74%	37.74%
HPVepi_T-cell (MHC-II)	130	67.19%	69.70%	68.46%

SEN, sensitivity; SPEC, specificity; ACC, accuracy

HPVepi web server

HPVepi algorithm for the prediction of HPV B-cell and T-cell (MHC-I and II) epitopes is integrated and implemented on the HPVomics resource. It is freely available at http://http://bioinfo.imtech.res.in/manojk/hpvomics/hpvepi.php. It is developed using PERL, PHP and SVM^{light} package. In HPVepi, user asked to enter or upload peptide sequences in FASTA format. We have also provided example sequences for guidance. Peptide sequences were further subjected for all three predictive models i.e. for B-cell, T-cell (MHC-I) and T-cell (MHC-II) to predict the peptide potentiality as epitope for all three arms of immunity.

On the web server, user has to provide peptide sequences of any length. We have developed separate methods for all three epitope classes. The output displays peptide name, sequences, B-cell score and prediction outcome, T-cell (MHC-I) score and result, T-cell (MHC-II) score and outcome, and most importantly potentiality of a peptide. Input and output of the HPVepi web server is shown in **Figure 53** using example sequences.

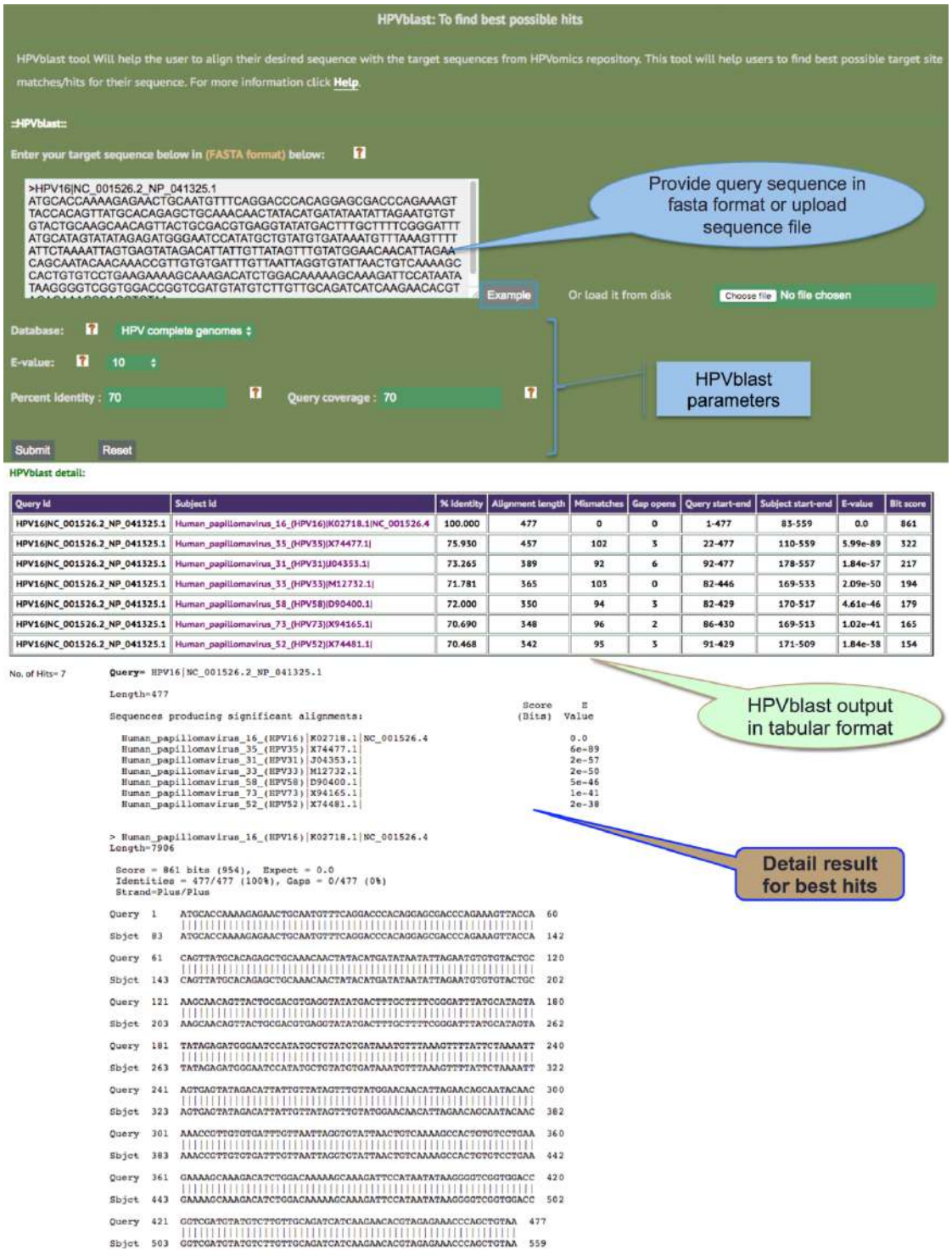


Name	Peptide sequences	B-score	Prediction Results	MI-score	Prediction Results	MII-score	Prediction Results	Potentiality
13BN	GGLGIGTGSGTGG	0.1	B-cell Epitope	0.6	MHC-I Epitope	0.2	MHC-II Epitope	★★★
14M2P	DSTLRLCVQSTHVD	0.7	B-cell Epitope	-0.9	Non-epitope	0.9	MHC-II Epitope	★★
15BN	IRYSRIGNKQTLRTR	-1	Non-epitope	0	MHC-I Epitope	0.6	MHC-II Epitope	★★
16M2N	ENVPDDLYIKGSGSTA	0.9	B-cell Epitope	0.9	MHC-I Epitope	-0.2	Non-epitope	★★
18M1N	QLLLGTLNIVCPICAPKT	1	B-cell Epitope	-0.8	Non-epitope	-0.2	Non-epitope	★
20BP	MHGDTPTLHEYMLDLQPETT	1	B-cell Epitope	0.7	MHC-I Epitope	-1	Non-epitope	★★
20M2N	VPPPNPVSKVVATDAYVTRT	-1.2	Non-epitope	-0.1	Non-epitope	-1	Non-epitope	--
8M1N	DIRTLEDL	0.7	B-cell Epitope	-0.8	Non-epitope	0.6	MHC-II Epitope	★★
9M1P	FAFRDLCIV	0.3	[illegible]	1.5	MHC-I Epitope	1	MHC-II Epitope	★★★

Figure 53. Screenshot representing input and output of HPVepi web server

Analysis tools

HPVomics also offer different analysis tools to explore genomic and proteomic content. These are as follows (1) HPVblast: this tool allow user to align query sequence to the HPV genomes and genes. The output of this tool is in tabular as well as in detail format (**Figure 54**). (2) ConBlock: this tool can be explored to select DNA and protein conserved and variable region from MSA (**Figure 55**). Here, Gblocks program (Castresana, 2000) was implemented. (3) Physicoprop: a significant tool to explore physico-chemical properties of peptides or epitopes.



Query id	Subject id	% identity	Alignment length	Mismatches	Gap opens	Query start-end	Subject start-end	E-value	Bit score
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_16_(HPV16)|K02718.1|NC_001526.4	100.000	477	0	0	1-477	83-559	0.0	861
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_35_(HPV35)|X74477.1|	75.930	457	102	3	22-477	110-559	5.99e-89	322
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_31_(HPV31)|J04353.1|	73.265	389	92	6	92-477	178-557	1.84e-57	217
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_33_(HPV33)|M12732.1|	71.781	365	103	0	82-446	169-533	2.09e-50	194
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_58_(HPV58)|D90400.1|	72.000	350	94	3	82-429	170-517	4.61e-46	179
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_73_(HPV73)|X94165.1|	70.690	348	96	2	86-430	169-513	1.02e-41	165
HPV16|NC_001526.2_NP_041325.1	Human_papillomavirus_52_(HPV52)|X74481.1|	70.468	342	95	3	91-429	171-509	1.84e-38	154

```
No. of Hits= 7

Query= HPV16|NC_001526.2_NP_041325.1

Length=477

Sequences producing significant alignments:                          Score     E
                                                                     (Bits)  Value

  Human_papillomavirus_16_(HPV16)|K02718.1|NC_001526.4                 0.0
  Human_papillomavirus_35_(HPV35)|X74477.1|                            6e-89
  Human_papillomavirus_31_(HPV31)|J04353.1|                            2e-57
  Human_papillomavirus_33_(HPV33)|M12732.1|                            2e-50
  Human_papillomavirus_58_(HPV58)|D90400.1|                            5e-46
  Human_papillomavirus_73_(HPV73)|X94165.1|                            1e-41
  Human_papillomavirus_52_(HPV52)|X74481.1|                            2e-38


> Human_papillomavirus_16_(HPV16)|K02718.1|NC_001526.4
Length=7906

 Score = 861 bits (954),  Expect = 0.0
 Identities = 477/477 (100%), Gaps = 0/477 (0%)
 Strand=Plus/Plus

Query  1    ATGCACCAAAAGAGAACTGCAATGTTTCAGGACCCACAGGAGCGACCCAGAAAGTTACCA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  83   ATGCACCAAAAGAGAACTGCAATGTTTCAGGACCCACAGGAGCGACCCAGAAAGTTACCA  142

Query  61   CAGTTATGCACAGAGCTGCAAACAACTATACATGATATAATATTAGAATGTGTGTACTGC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  143  CAGTTATGCACAGAGCTGCAAACAACTATACATGATATAATATTAGAATGTGTGTACTGC  202

Query  121  AAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  203  AAGCAACAGTTACTGCGACGTGAGGTATATGACTTTGCTTTTCGGGATTTATGCATAGTA  262

Query  181  TATAGAGATGGGAATCCATATGCTGTATGTGATAAATGTTTAAAGTTTTATTCTAAAATT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  263  TATAGAGATGGGAATCCATATGCTGTATGTGATAAATGTTTAAAGTTTTATTCTAAAATT  322

Query  241  AGTGAGTATAGACATTATTGTTATAGTTTGTATGGAACAACATTAGAACAGCAATACAAC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  323  AGTGAGTATAGACATTATTGTTATAGTTTGTATGGAACAACATTAGAACAGCAATACAAC  382

Query  301  AAACCGTTGTGTGATTTGTTAATTAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAA  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  383  AAACCGTTGTGTGATTTGTTAATTAGGTGTATTAACTGTCAAAAGCCACTGTGTCCTGAA  442

Query  361  GAAAAGCAAAGACATCTGGACAAAAAGCAAAGATTCCATAATATAAGGGGTCGGTGGACC  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  443  GAAAAGCAAAGACATCTGGACAAAAAGCAAAGATTCCATAATATAAGGGGTCGGTGGACC  502

Query  421  GGTCGATGTATGTCTTGTTGCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAA  477
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  503  GGTCGATGTATGTCTTGTTGCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAA  559
```

Figure 54. Screenshot representing HPVblast tool and output in tabular along with mapping format

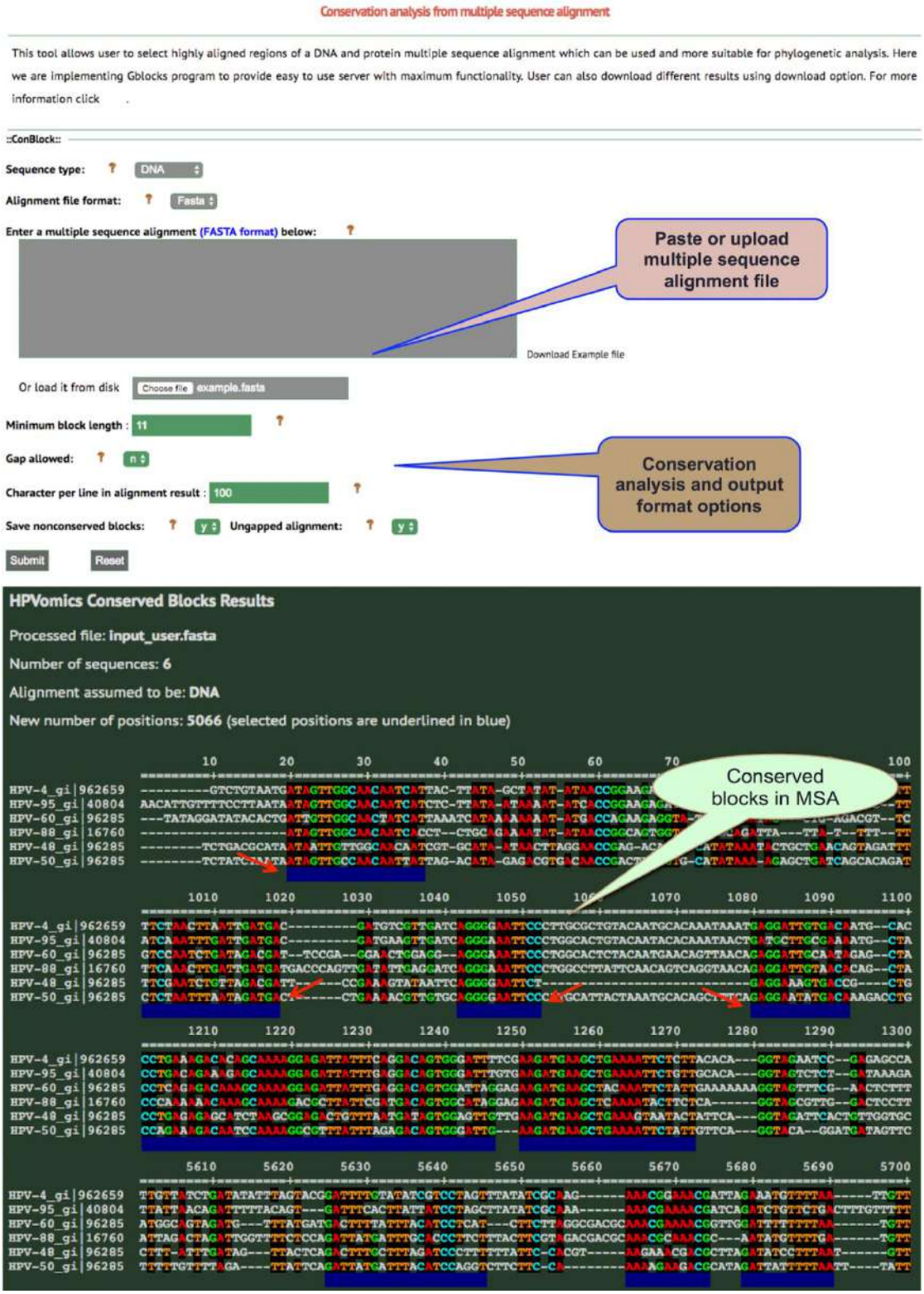


Figure 55. ConBlock output depicting conserved block regions in multiple sequence alignment

Conclusion

HPV associated carcinogenesis is a global health problem and multi therapeutic strategy may open new ways to expedite design and development of effective combat strategy against oncogenic HPVs along with regular screening and vaccination. In the present study, efforts are made to perform systematic assessment for the identification and analysis of putative therapeutic regimens targeting different HPV genes and proteins specially oncogenes. We have developed a user-friendly web resource "HPVomics", which delivers therapeutically important elements such as vaccine epitope candidates (HPV epitome), anti-viral peptides, RNA based solutions and pathway information. It provides a blend of potential therapeutic knowledge, epitome, interactive genomic annotation browser and genomic analysis, which craft it for broader research applications. It also encompasses first HPV specific epitome prediction method, i.e., HPVepi. Up to now, there are few in-silico attempts are made that focus on HPVs. However, there is no such resource or compendium available. We anticipate that, this resource will be useful for wider research community with special focus on HPV epitopes and therapeutics.

Benchmarking of de novo genome assemblers for the viral next generation sequencing (NGS) data

Chapter 5. Benchmarking of *de novo* genome assemblers for the viral next generation sequencing (NGS) data

Introduction

Next generation sequencing (NGS) proven to be valuable in the field of virology. Various studies have shown the applications of NGS in viral research including diagnostics. This broadly cover virus identification and diversity (Barzon et al., 2011a; Briese et al., 2015; Capobianchi et al., 2013; Foulongne et al., 2012a; Hannigan et al., 2015; Lecuit and Eloit, 2013; Scarpellini et al., 2015; Wylie et al., 2013).

One of the most considered and a crucial step in NGS data analysis is the genome assembly. This is the process to generate large contigs from the raw small reads. With the advancement in sequencing technologies, various de novo genome assembly software tools based on different algorithms were developed (de Freitas et al., 2014). Mainly, overlap layout consensus (OLC) and *de Bruijn* graph (dBg) based assemblers were established (Wajid and Serpedin, 2012). This contains some of the widely employed assemblers such as Velvet (Zerbino and Birney, 2008), Edena (Hernandez et al., 2008), SOAPdenovo (Li et al., 2010b), ABySS (Assembly By Short Sequences), IDBA (Iterative De Bruijn graph Assembler) (Peng et al., 2012), SPAdes (Bankevich et al., 2012), ALLPATHS-LG etc. (Miller et al., 2010). Along with this, some virus specific assembly tools, i.e., IVA (Hunt et al., 2015), VICUNA (Yang et al., 2012), VGA (Mangul et al., 2014), Arapan-S (Sahli and Shibuya, 2012), etc. were also developed.

There are distinct studies regarding the evaluation and comparison of genome assembly algorithms based on different measures (Bao et al., 2011; Earl et al., 2011; Finotello et al., 2012; Magoc et al., 2013; Salzberg et al., 2012; Zhang et al., 2011). These evaluations mainly focused on and utilized the data from human, bacterial or plant origins (Barthelson et al., 2011; Earl et al., 2011; Salzberg et al., 2012). None of these broad assessment studies have shown the performance of existing assembly algorithms on viral NGS data. Therefore, there is a need for the comparison and benchmarking of different assemblers on viral raw sequencing data from different sequencing platforms.

Materials and Method

For the benchmarking of distinct assembly tools on viral NGS data, number of steps is followed. These primarily include installation and configuration, raw data collection, quality control and evaluation, genome assembly and assessment. The complete workflow is depicted in **Figure 56**.

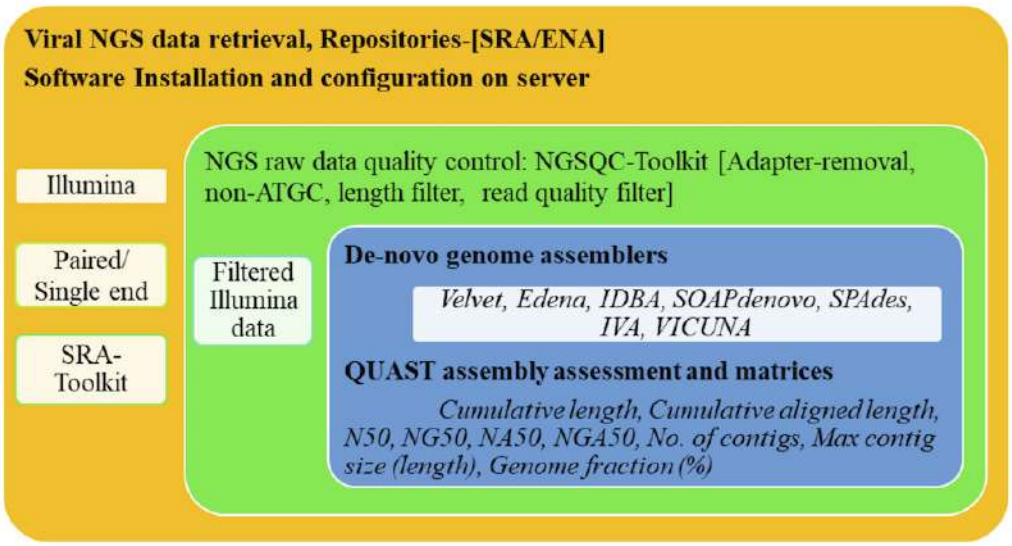


Figure 56. Diagram showing outline of methodology used in the study

Data retrieval

Viral NGS raw data from different Illumina platforms were retrieved from the freely available repositories namely European Nucleotide Archive (ENA)(https://www.ebi.ac.uk/ena) and Sequence Read Archive (SRA)(https://www.ncbi.nlm.nih.gov/sra). Paired or single end read data were extracted. Further, summarized statistics table was prepared for all the data.

Installation and configuration

Different tools for quality control, assembly and analysis were installed and configured on the server (operating system (OS): Ubuntu, RAM: 528GB, CPUs: 160, Threads/core: 2) without root permissions. This include NGSQC Toolkit v2.3 for the quality control (QC) of data, assembly tools Velvet-1.2.09, SOAPdenovo-127mer-v2.04, Edena v3-131028, IDBA-1.1.0, ABySS-1.9.0, SPAdes, IVA and VICUNA and quality assessment tool QUAST v2.2.

Quality control and evaluation

Raw NGS data were first subjected to quality analysis to remove and filter out various sequencing errors, i.e., read trimming (primer/adapter/low complexity reads), homopolymer trimming, length filtering, contamination removal, etc. utilizing NGSQC-Toolkit for both Illumina read data (**Table 20**). Quality filtered data were evaluated and various number statistics were listed (**Table 21**).

Table 20. Parameters used for the quality control

Parameter name	Values
Primer/Adaptor library	Genomic DNA/Chip-seq Library
Cut-off read length for HQ	70%
Cut-off quality score	30

Genome assembly and assessment

After quality control, high quality reads were subjected to assembly for the reconstruction of viral genome. Eight different assemblers were used on different real NGS data from diverse viruses i.e. ssRNA (+ve) virus, ssRNA (-ve) virus, retro-transcribing virus, dsRNA virus and dsDNA virus. We have developed distinct in-house scripts to automate the processes in order to perform assembly on diverse set of k-mers and overlaps for different assemblers.

Genome assembly using SOAPdenovo

SOAPdenovo is a dBg-based assembly tool works with Illumina data. It mainly works in two steps, pregraph and contig. We have generated assemblies for each data utilizing minimum k-mer of 23 to either maximum 127 or the length of read with step of 2. General commands other than defaults employed are denoted below.

SOAPdenovo-127mer pregraph -s <configFile> -o <output-directory/filename> -K <k-mer> -p <cpu_number>

SOAPdenovo-127mer contig -g <output-directory/filename> -p <cpu_number>

Genome assembly using Velvet

Velvet is a dBg-based denovo assembler. It comprises of two main programs. First, velveth (hashing program) and second velvetg (de Bruijn graph construction, error removal and repeat resolution). Assemblies were generated for all the k-mers from 23 (min-hash) to 59 (max-hash), with step of 2. General commands used are as follows:

velveth output-directory/ <min-hash, max-hash, step> -fastq -separate -shortPaired <input-file> (forward and reverse reads)

#For short read data (GA II, GA IIx, Hiseq 2000)

velveth output-directory/ <min-hash, max-hash, step> -fastq -separate -longPaired <input-file> (forward and reverse reads)

#For long read data (Miseq, Nextseq)

velvetg output-directory/ -cov_cutoff auto -ins_length <int> -min_contig_lgth <int> -amos_file yes -exp_cov auto -scaffolding yes -unused_reads yes

Genome assembly using ABySS

ABySS is a dBg-based parallelized short sequence assembler. It uses number of inbuilt modules to perform different task to generate unitig, contigs and scaffolds. Abyss-fac can also calculate and provide assembly statistics. Assemblies with minimum 23 k-mer to maximum 128 or length of reads were produced with step of 5.

Set of commands utilized are

abyss-pe name=<filename-prefix> k=<k-mer size> G=<genome-size> n=<minimum number of pairs to build contig 2 is used> s=<minimum unitig size 100 used> S=<minimum contig size 100 used> l=<minimum alignment length equal to k-mer used> in='<inputfiles (forward and reverse)>' #For paired end Illumina data

abyss-pe name=<filename-prefix> k=<k-mer size> G=<genome-size> n=<2> s=<100> S=<n> l=<k-mer_size> se='<inputfile>' #For single end Illumina data

Genome assembly using IDBA

IDBA is an iterative dBg-based assembly tool having different mode for short and long read data. Set of commands implemented is as follows.

fq2fa --paired <forward and reverse input file><directory-name/filename.fa>

idba_pe -r <directory-name/filename.fa> --num_threads <no. of threads> -o <directory-name> #Short mode (read length <100)

idba_pe -l <directory-name/filename.fa> --num_threads <no. of threads> --mink <minimum k-mer> -o <directory-name> #Long mode (read length >100)

Genome assembly using SPAdes

SPAdes-St. Petersburg genome assembler is a dBg-based tool for both single-cell and multicell data assemblies. It supports Illumina paired-end, mate-pairs and unpaired reads.

Set of commands utilized are

spades.py -k <kmer values> -1 <input-read-file-forward> -2 <input-read-file-reverse> -o <output-directory>

Genome assembly using Edena

Edena is an OLC based algorithm. It works in two modes, i.e., overlapping and assembly. Used commands except default parameters are mentioned below:

edena -nThreads <no. of threads> -DRpairs <input-read-files-forward, reverse> -M <minimum size of overlap> -p <output-directory/filename> #Overlap mode

edena -e <output-directory/filename.ovl> -c <minimum contig length> -p <output-directory/filename1> #Assembly mode

Genome assembly using IVA

Iterative Virus Assembler (IVA) is designed specifically for read pairs sequenced at highly variable depth from RNA virus samples. Command utilized is

iva --max_contigs <number> -f <input-read-file-forward> -r <input-read-file-reverse><directory>

Genome assembly using VICUNA

Vicuna is an OLC based de novo assembly tool that generates consensus assemblies from heterogenous and diverse viral population data.

For this a config file is provided with the different parameters for assembly like pFqDir: input directory for paired fastq files, npFqDir: input directory for non-paired fastq files, batchSize, min_output_contig_len: minimum length of contigs, outputDIR: output directory. Other parameters were taken as default.

Genome assembly assessment and comparison using QUAST

Subsequently, assessment and comparison of different assemblies is performed based on discrete methods and criteria mainly assembly lengths, N50, NG50, NA50,

NGA50, contig length (largest), N50, number of contigs, and genome fraction percentage etc. utilizing quality analysis tool QUAST. General command used is:

quast.py -o <output-directory-name> --min-contig <minimum contig length> -f -S 100,200,300,500,1000 --est-ref-size <reference size> -t 100,200,300,500,1000 -s <input contigs file>

General command used for the plotting and comparing different assemblies for a particular viral data is as follows:

quast.py -o <output-directory-name> -R <reference.fasta> -G <reference.gff3> --min-contig <minimum contig length> -l <comma-seprated-assembly-lables> -f -S 100,200,300,500,1000 --gage -t 100,200,300,500,1000 -s <input contigs fasta files according to assembly labels>

Results and discussion

In the study, we are reporting the quality control, analysis and evaluation of genome assemblies on the viral NGS datasets (paired and single end) of different viral categories, i.e., ssRNA (+ve) viruses (like Dengue virus (DENV), West Nile virus (WNV),Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)), ssRNA (-ve) virus (Influenza viruses), retro-transcribing viruses (like Human immunodeficiency virus (HIV), Hepatitis B virus (HBV)), dsDNA viruses (Human herpesvirus (HHV), Human papillomavirus (HPV)) using widely employed assemblers, i.e., SOAPdenovo, Velvet, ABySS, IDBA, SPAdes, Edena, IVA and VICUNA. Here, we are also considering the different sequencing platforms of Illumina (GA II, GA IIx, HiSeq 2000, MiSeq, Nextseq 500) (**Table 21**).

Viral NGS raw data and quality control

Overall, nine viral NGS data set of Illumina sequencing platforms (GAII, GAIIx, Hiseq, Miseq, Nextseq 500) were retrieved. Summarized statistics table were prepared for the data (**Table 21**). Different paired and single end reads in fastq format were subjected to quality control and filtering. For QC inbuilt genomic DNA primer/adaptor library, high quality (HQ) read length cut-off (70%) and quality score cut-off (30) is used to remove contamination and low-quality reads. QC statistics for all the 9 datasets listed in **Table 21** is depicted through distinct graphs for each data set (**Figure 57-65**).

Table 21. The Illumina viral NGS data and quality analysis statistics

Viruses	Run Accession	Technology	Layout	Number of raw reads	Number of quality filtered reads	Reads length (*bp*)
Influenza virus A	*ERR045841*	*GA II*	*Paired*	*5,18,134*	*3,66,219*	*54*
HHV 8	*ERR244026*	*GA IIx*	*Paired*	*2,955,212*	*1,784,389*	*76*
HIV 1	*SRR527726*	*HiSeq 2000*	*Paired*	*3,488,150*	*2,435,004*	*101*
Rhinovirus A	*SRR499802*	*HiSeq 2000*	*Paired*	*16,947*	*7,493*	*101*
DENV 3	*SRR546416*	*MiSeq*	*Paired*	*4,72,546*	*15,800*	*225*
WNV	*SRR546546*	*MiSeq*	*Paired*	*1,61,067*	*881*	*225*
HBV	*DRR001353*	*GA IIx*	*Single*	*7,68,941*	*4,50,365*	*64*
HPV-16	*SRR8607785*	*NextSeq 500 500*	*Paired*	*161706*	*157592*	*149-151*
SARS-CoV-2	*SRR11597222*	*MiSeq*	*Paired*	*108214*	*93046*	*292-301*

From Influenza virus A (ERR045841) Illumina GAII paired-end data raw reads of length 54 bp are 5,18,134, after quality control 3,66, 219 high quality reads (~71%) were remained and ~29% were removed (**Figure 57**). From HHV 8 Illumina (ERR244026) GAIIx data with the 2,955,212 raw reads (76 bp) were quality filtered and quality filtered reads are 1,784,389 (~60%). The remaining low-quality reads (~40%) were discarded (**Figure 58**). Further, two datasets belong to the HiSeq platform. HIV 1 (SRR527726) having the 3,488,150 raw reads of length 101 bp and retained quality filtered reads are 2,435,004 (~70%) (**Figure 59**). Rhinovirus A (SRR499802) having 16,947 reads (101 bp) were quality filtered and quality reads are 7,493 (~44%) (**Figure 60**). Further, DENV 3 MiSeq (SRR546416) data consist of 4,72,546 raw reads and 15,800 (3.5%) quality filtered reads of length 225 bp (**Figure 61**). Similarly, WNV MiSeq (SRR546546) raw data is having 1,61,067 reads (225 bp) and the quality filtered reads are 881 (~1%) (**Figure 62**). Further, HBV GA IIx single-end (DRR001353) data of 7,68,941 reads quality filtered and 4,50,365 (~58%) quality reads were obtained (**Figure 63**). Moreover, we have also included HPV-16 and SARS-CoV-2 NGS data in benchmark. HPV-16 NextSeq 500 (SRR8607785) data consist of 1,61,706 raw reads (149-151 bp) and 1,57,592 (~97) quality filtered reads were found (**Figure 64**). Likewise, SARS-CoV-2 MiSeq (SRR11597222) data of length 292-301 having 108214 reads quality filtered. Total, 93046 (~86%) quality reads were retained (**Figure 65**).

Genome assembly and assessment

Next, we have performed the genome assembly utilizing different quality filtered data from diverse platforms as previously mentioned.

Illumina data analysis is done using different assemblers namely SOAPdenovo, Velvet, ABySS, IDBA, SPAdes, Edena, IVA, and VICUNA. We have generated the assemblies at different k-mers or overlaps to evaluate and obtain best likely assembly result. Different parameters viz. cumulative assembly length, cumulative aligned length, N50, NG50, NA50, NGA50, largest contig, number of contigs, genome fraction percentage were evaluated from each assembler to deduce comprehend picture (**Table 22 and Figures 57-65**).

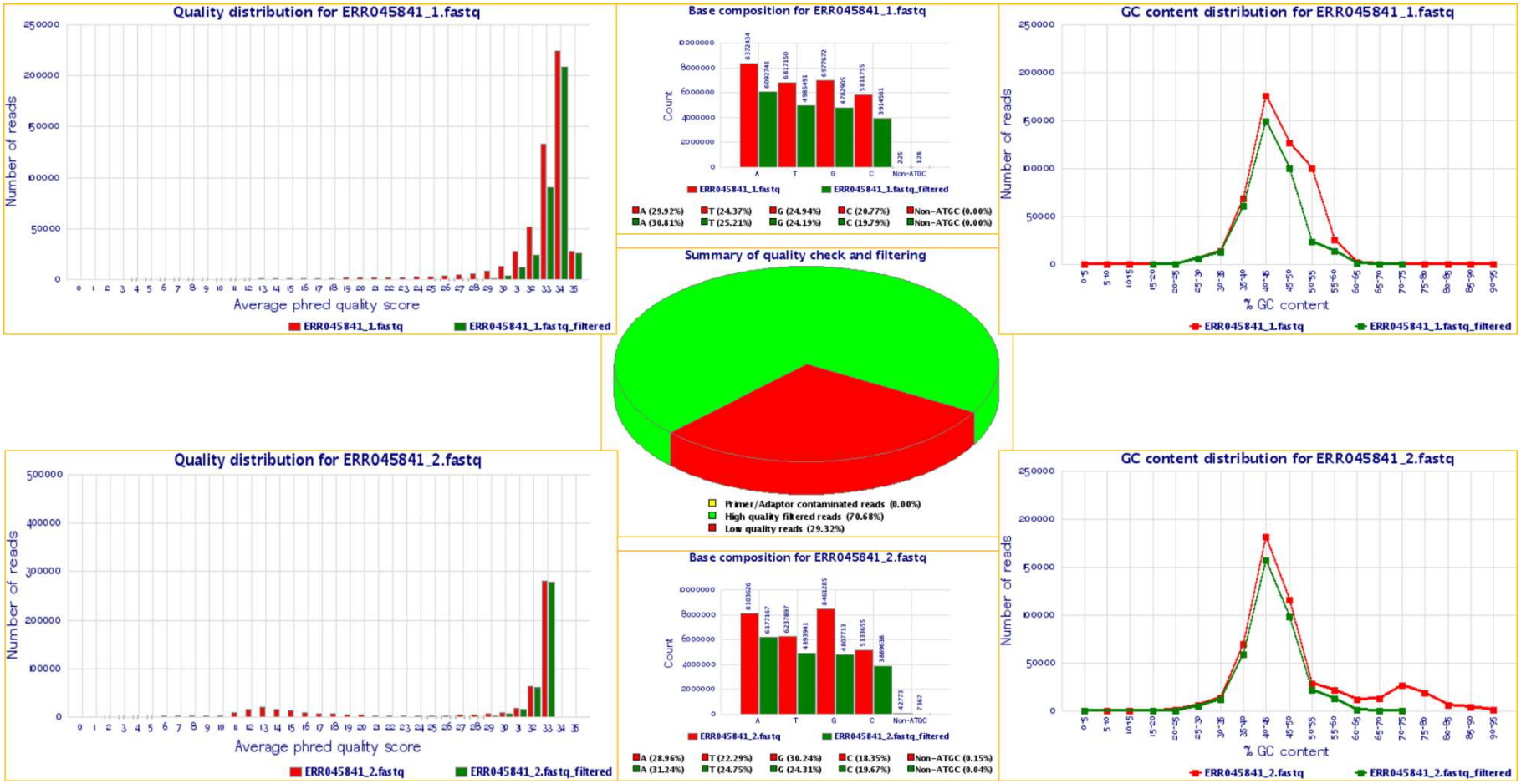


Figure 57. Quality control statistics of Influenza virus A (ERR045841)

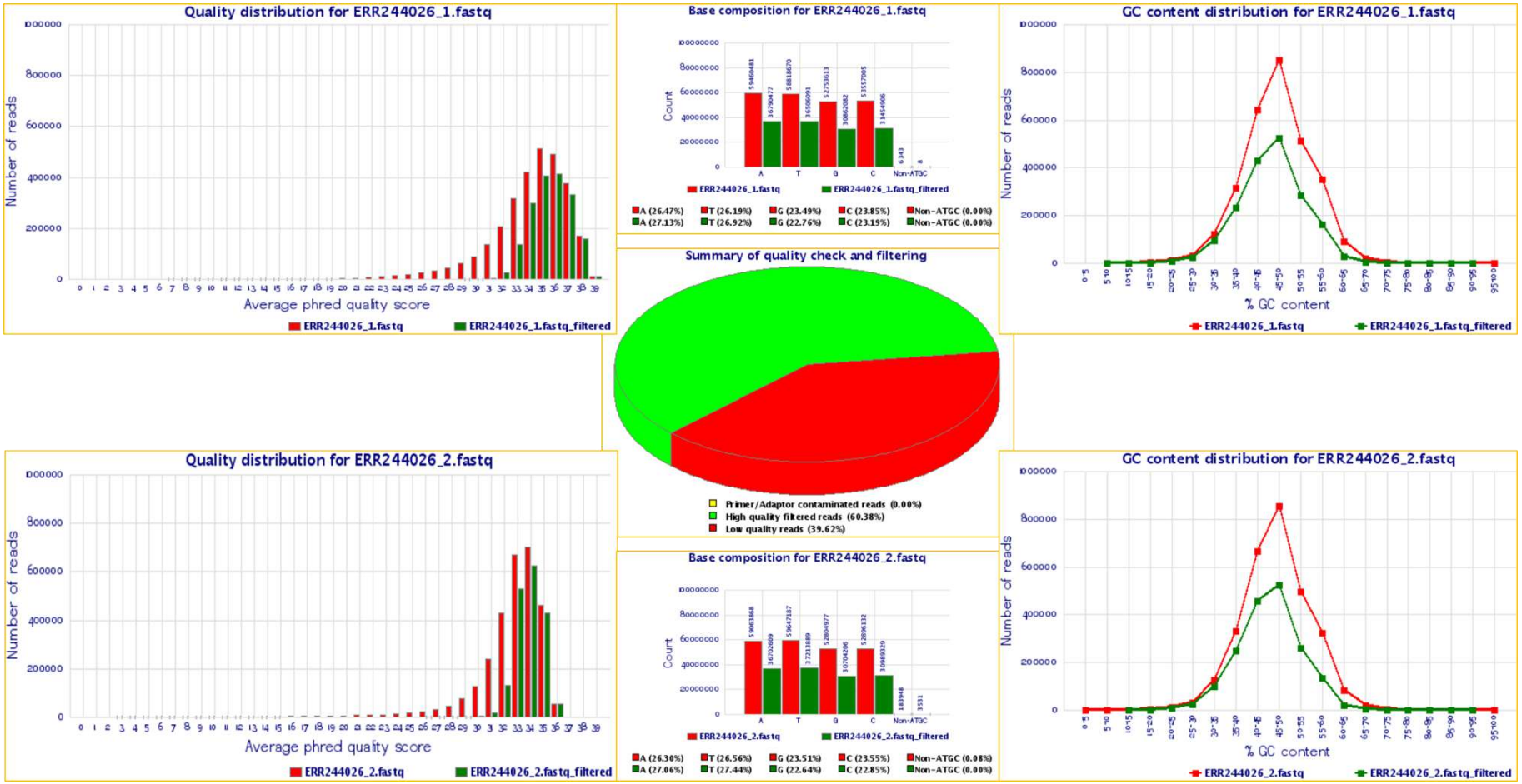


Figure 58. Quality control statistics of Human herpesvirus 8 (HHV 8) (ERR244026)

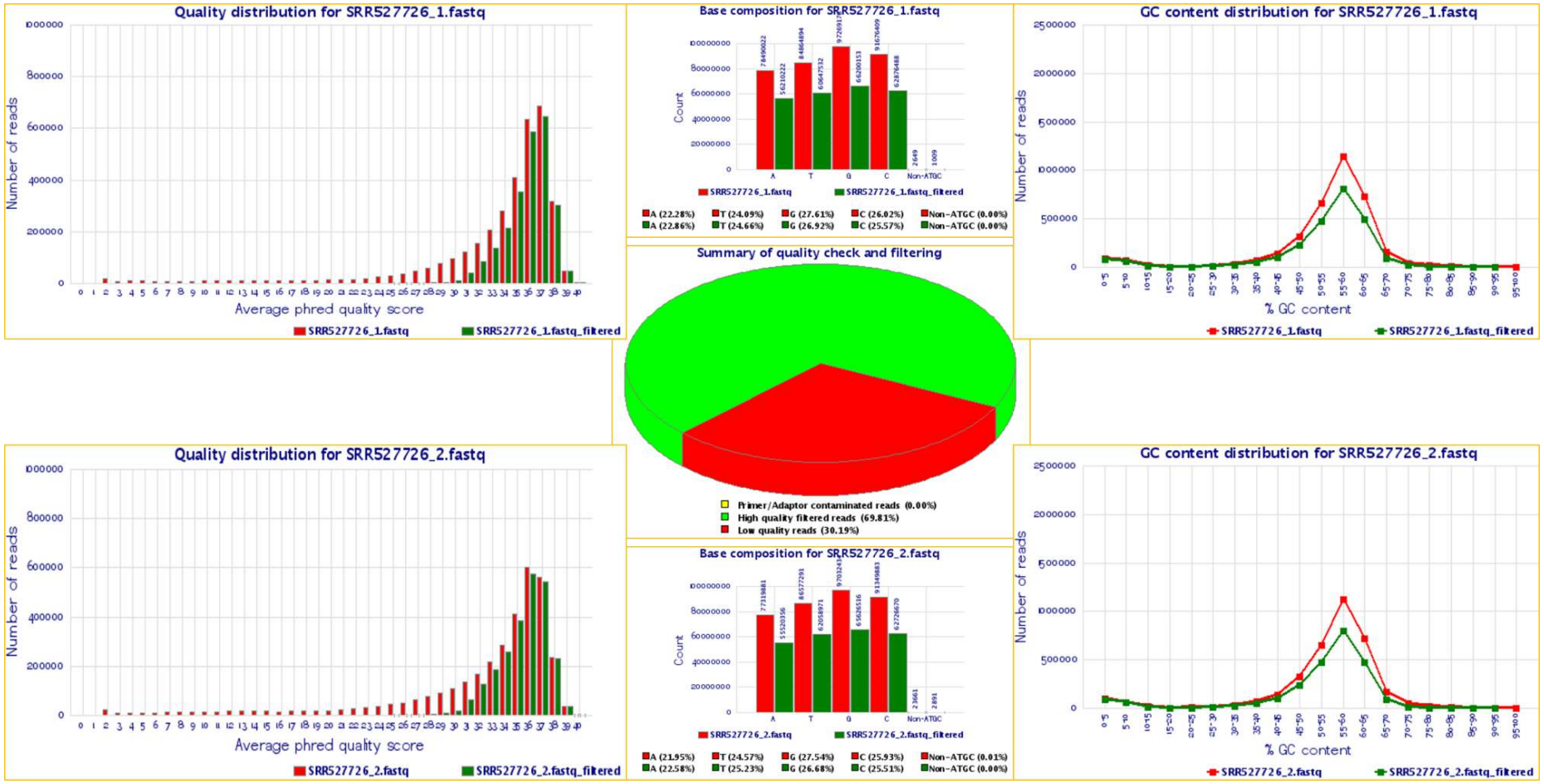


Figure 59. Quality control statistics of Human immunodeficiency virus 1 (HIV 1) (SRR527726)

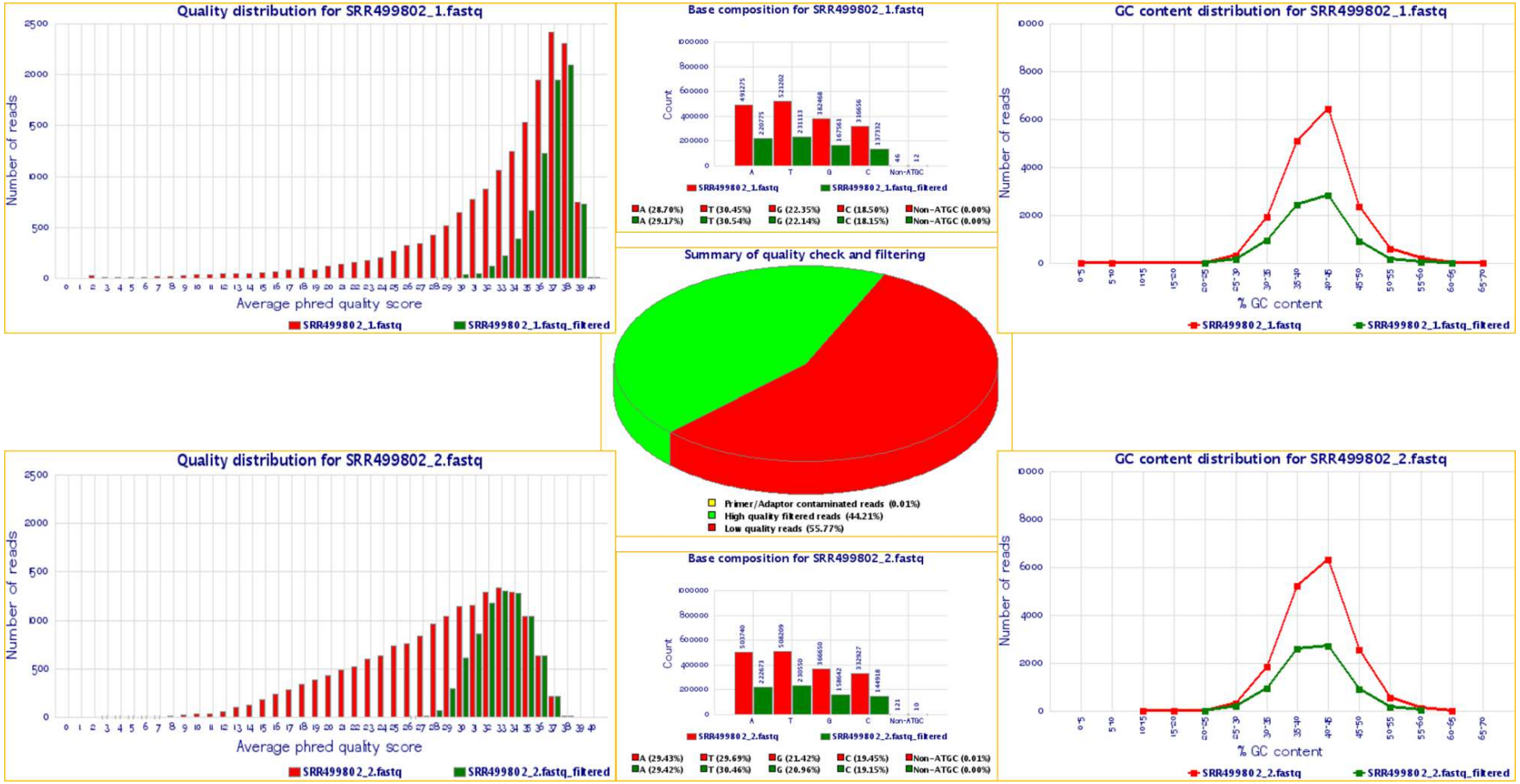


Figure 60. Quality control statistics of Rhinovirus A (SRR499802)

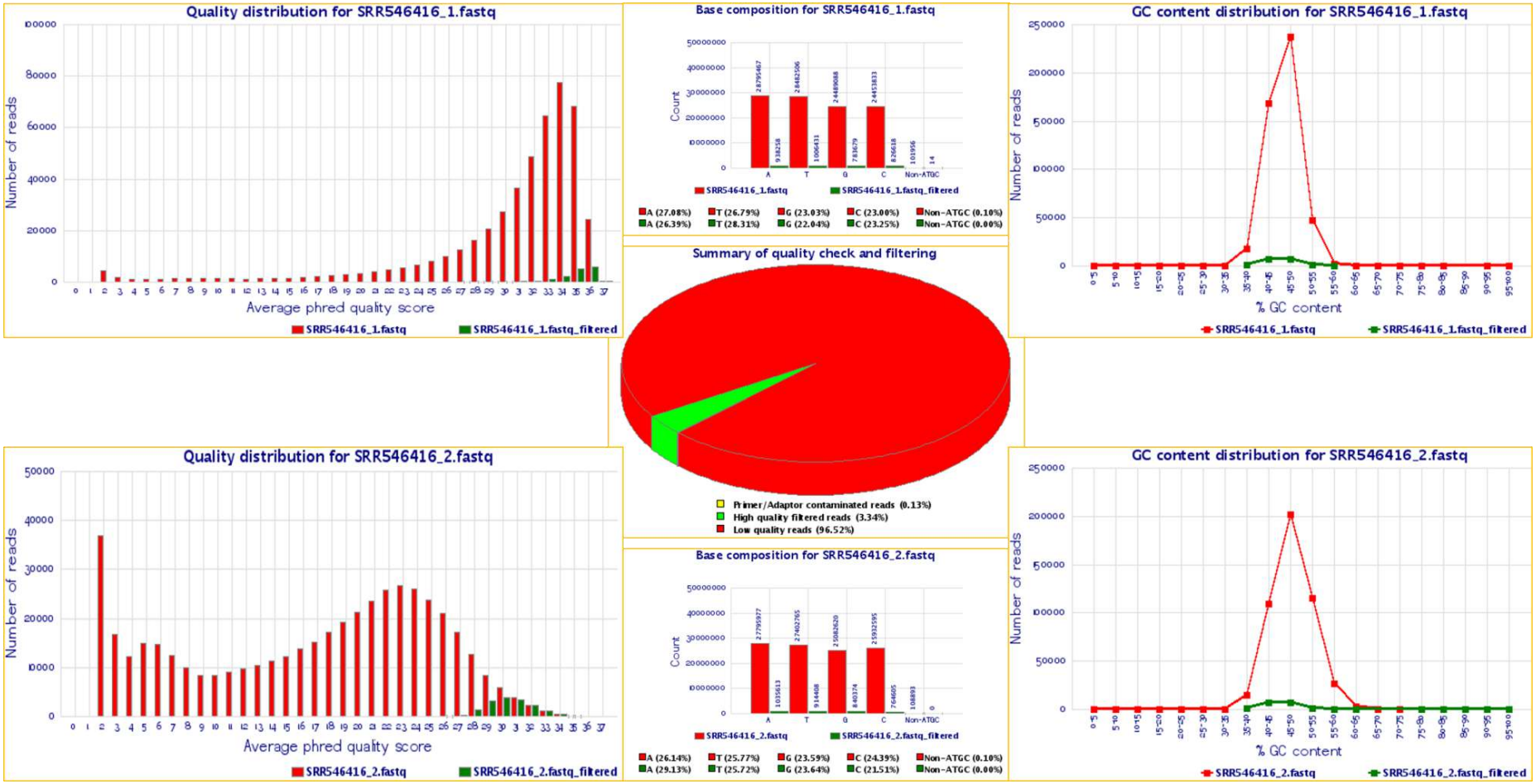


Figure 61. Quality control statistics of Dengue virus 3 (DENV 3) (SRR546416)

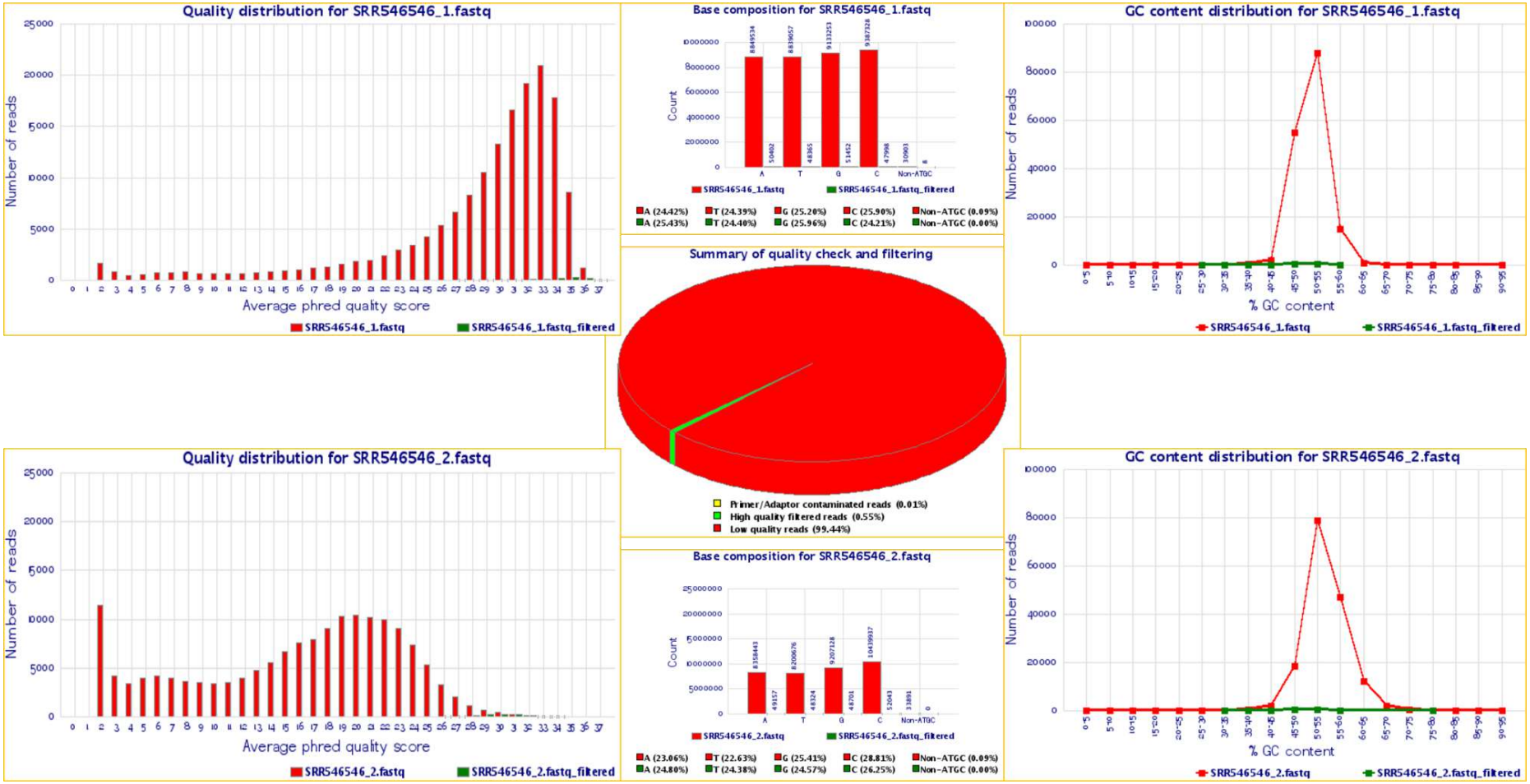


Figure 62. Quality control statistics of West Nile virus (WNV) (SRR546546)

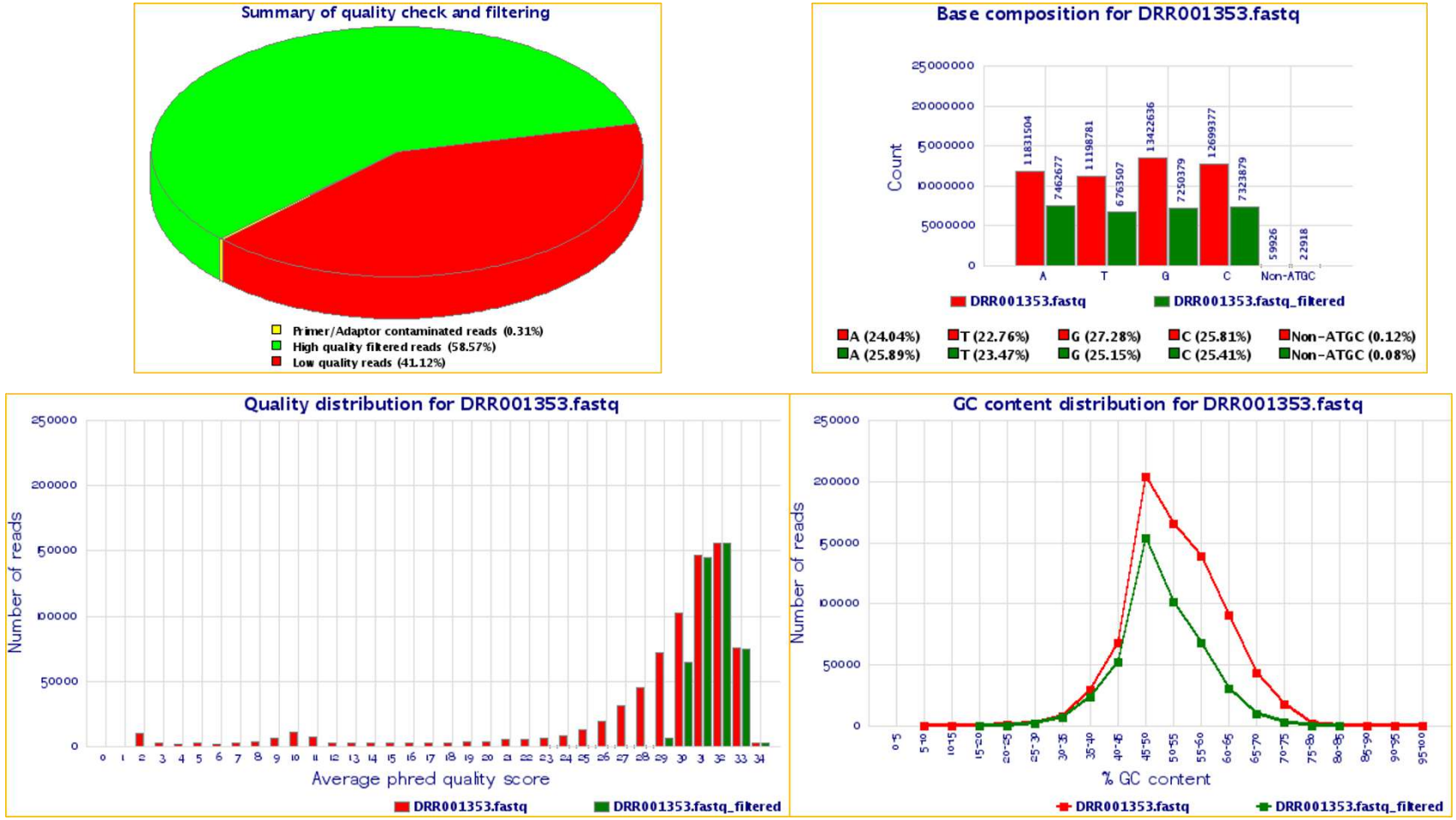


Figure 63. Quality control statistics of Hepatitis B virus (HBV) (DRR001353)

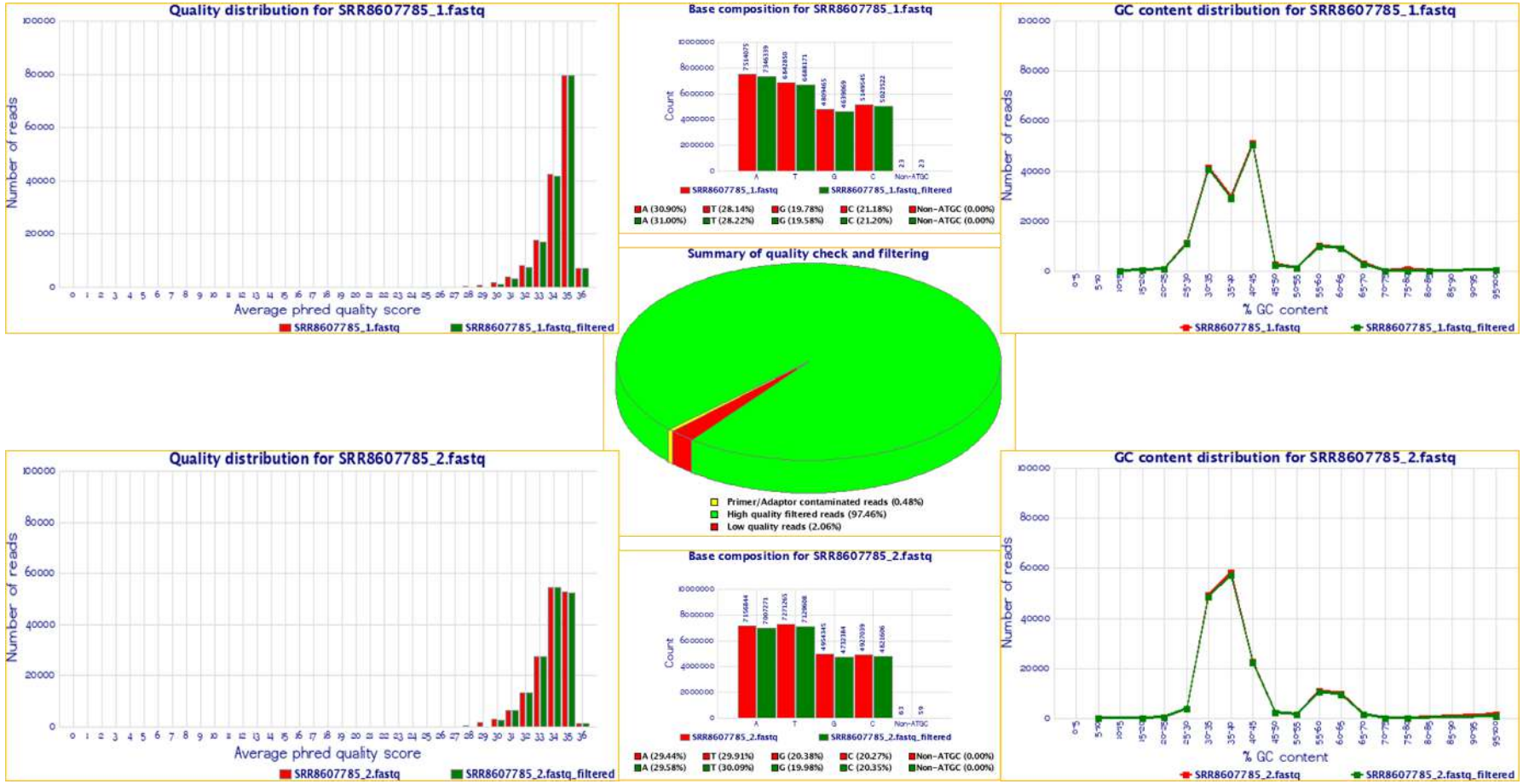


Figure 64. Quality control statistics of Human papillomavirus 16 (HPV 16) (SRR8607785)

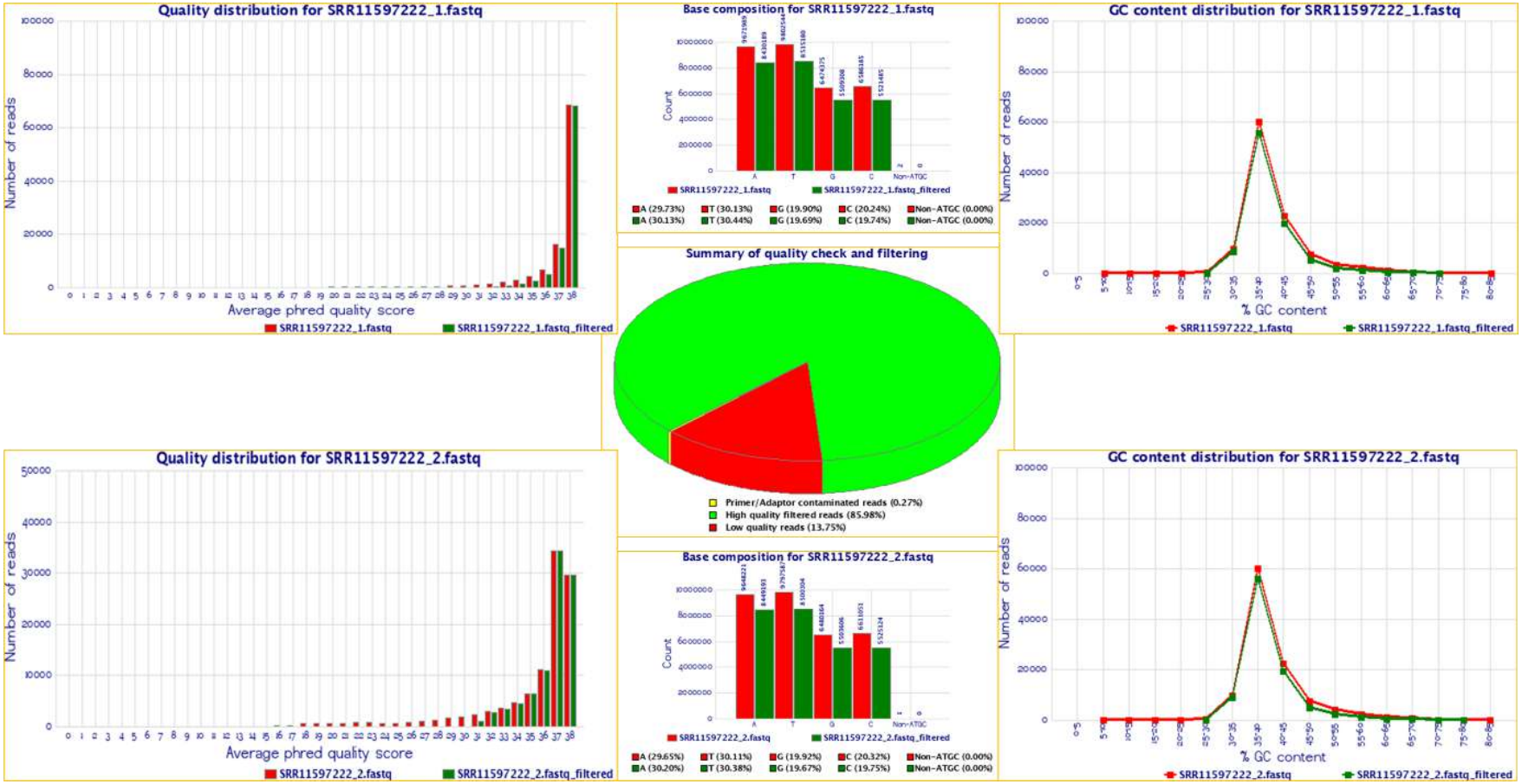


Figure 65. Quality control statistics of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (SRR11597222)

Table 22. Table illustrating genome fraction (coverage %), largest contig length distribution and N50 values from different assemblers for distinct viral Illumina data genome assemblies

METRICES	ASSEMBLERS	1	2	3	4	5	6	7	8	9
GENOME FRACTION (%)	SOAPdenovo/2.04	51.05	97.23	0	81.55	98.74	94.03	26.78	1.8	-
	Velvet/1.2.09	20.77	94.08	-	18.98	97.62	90.87	3.07	6.72	-
	ABySS	39.63	**99.98**	-	**95.6**	**100**	93.21	10.49	22.53	-
	IDBA/1.1.0	39.67	96.95	**70.59**	**95.6**	97.32	**94.6**	**77.85**	**92.25**	**78.29**
	SPAdes	**63.66**	**99.98**	11.83	**82.87**	**100**	**96.52**	**72.9**	**94.6**	3.49
	Edena/3.131028	37.16	96.94	0	74.45	99.42	59.46	0	0	-
	IVA	25.18	88.96	-	0	100	19.85	0	6.68	0
	VICUNA	**40.36**	**97.08**	**63.42**	**88.23**	**100**	**95.3**	**80.37**	0	0
LARGEST CONTIG LENGTH	SOAPdenovo/2.04	368	33854	0	2139	3475	3402	582	766	4052
	Velvet/1.2.09	1068	44860	1637	2231	2471	3937	833	1055	6245
	ABySS	725	**92945**	1718	**6648**	**10682**	**3947**	582	**1273**	4031
	IDBA/1.1.0	675	**62789**	**7052**	**3906**	2234	3328	**842**	**6375**	**4988**
	SPAdes	**1219**	**93048**	**9249**	**2289**	**7721**	**5865**	**1499**	**10482**	**4921**
	Edena/3.131028	869	29731	0	1411	5262	1565	0	0	4022
	IVA	1009	6302	3315	0	4334	884	0	699	0
	VICUNA	631	8952	**3775**	1170	**5280**	1361	**969**	0	0
N50	SOAPdenovo/2.04	196	9402	0	1644	3140	3254	273	538	1756
	Velvet/1.2.09	400	23877	1142	1598	2099	3355	263	560	1767
	ABySS	230	**92945**	1376	**6648**	**10682**	**3270**	**286**	**591**	**1939**
	IDBA/idba-1.1.0	177	**23914**	**2143**	**3906**	1912	2513	**381**	**808**	**2097**
	SPAdes	**446**	**93048**	**7741**	**1930**	**7721**	**5865**	**426**	584	**2420**
	Edena/v3.131028	243	17945	0	1046	2751	837	0	0	1765
	IVA	939	2280	2322	0	3583	689	0	673	0
	VICUNA	320	1787	1666	426	3136	594	279	0	0

1: Influenza_A_5841_GAII, 2: HHV_8_4026_GAIIx, 3:HIV_1_7726_Hiseq, 4: Rhinovirus_A_9802_HIseq, 5: DENV_3_6416_MIseq, 6: WNV_6546_Miseq, 7: HPV_16_7785_Nextseq, 8: SARS-CoV-2_7222_Miseq, 9: HBV_1353_GAIIx_single

Genome assembly of Influenza A virus (GAII) data (ERR045841)

We have performed genome assembly of influenza A data with the read length of 54 bp using all the mentioned tools. The reference used for the assembly comparison is CY116347. We are able to reconstruct and achieve genome fraction coverage percentage of maximum ~64% using SPAdes assembler and the least fraction (20%) is obtained from velvet assembler. The largest contig of length 1219 bp is also obtained from SPAdes tool. The performance of the distinct assemblers is in the following order **SPAdes > SOAPdenovo**> VICUNA > IDBA > ABySS > Edena >**IVA > Velvet** for the influenza A virus data. The detailed statistics and different performance matrices are shown in **Table 22** and **Figure 66**.

Genome assembly of HHV 8 (GAIIx) data (ERR244026)

We have performed genome assembly of HHV 8 data with the read length of 76 bp using all the mentioned tools. The reference used for the assessment is NC_009333. All the assemblers performed good on the HHV 8 data. However, the highest genome fraction coverage percentage (99.98%) is obtained through the SPAdes and ABySS tools based on the QUAST aligned statistics. The largest contig (maximum length) is obtained from the SPAdes of 93048 bp. Likewise; ABySS generated the largest contig of length 92945 bp. The least performed assembler on the HHV 8 GAIIx platform data is IVA with 88% genome fraction (%). The performance of these assemblers is in the following order **SPAdes**>**ABySS**> SOAPdenovo >VICUNA > IDBA> Edena >**Velvet > IVA** (**Table 22** and **Figure 67)**.

Genome assembly of HIV 1 (Hiseq) data (SRR527726)

Genome assembly of HIV 1 data of read length 101 bp is performed with different assemblers. Reference used for comparison of HIV-1 assemblies is FJ469707. However, two assemblers, i.e., SOAPdenovo and Edena were not able to generate any assembly. Likewise, other 3 assemblers Velvet, ABySS and IVA are not able to construct HIV genome. Only 3 assemblers are able to reconstruct HIV genome with the selected parameters. The maximum genome fraction percentage of ~71% is achieved by IDBA followed by VICUNA with ~63%. Likewise, SPAdes provide only genome fraction of ~12%. The largest contig length is obtained by SPAdes (9249 bp) and IDBA (7052 bp). The performance of the assemblers is in the following order **IDBA>VICUNA**>SPAdes (**Table 22** and **Figure 68)**.

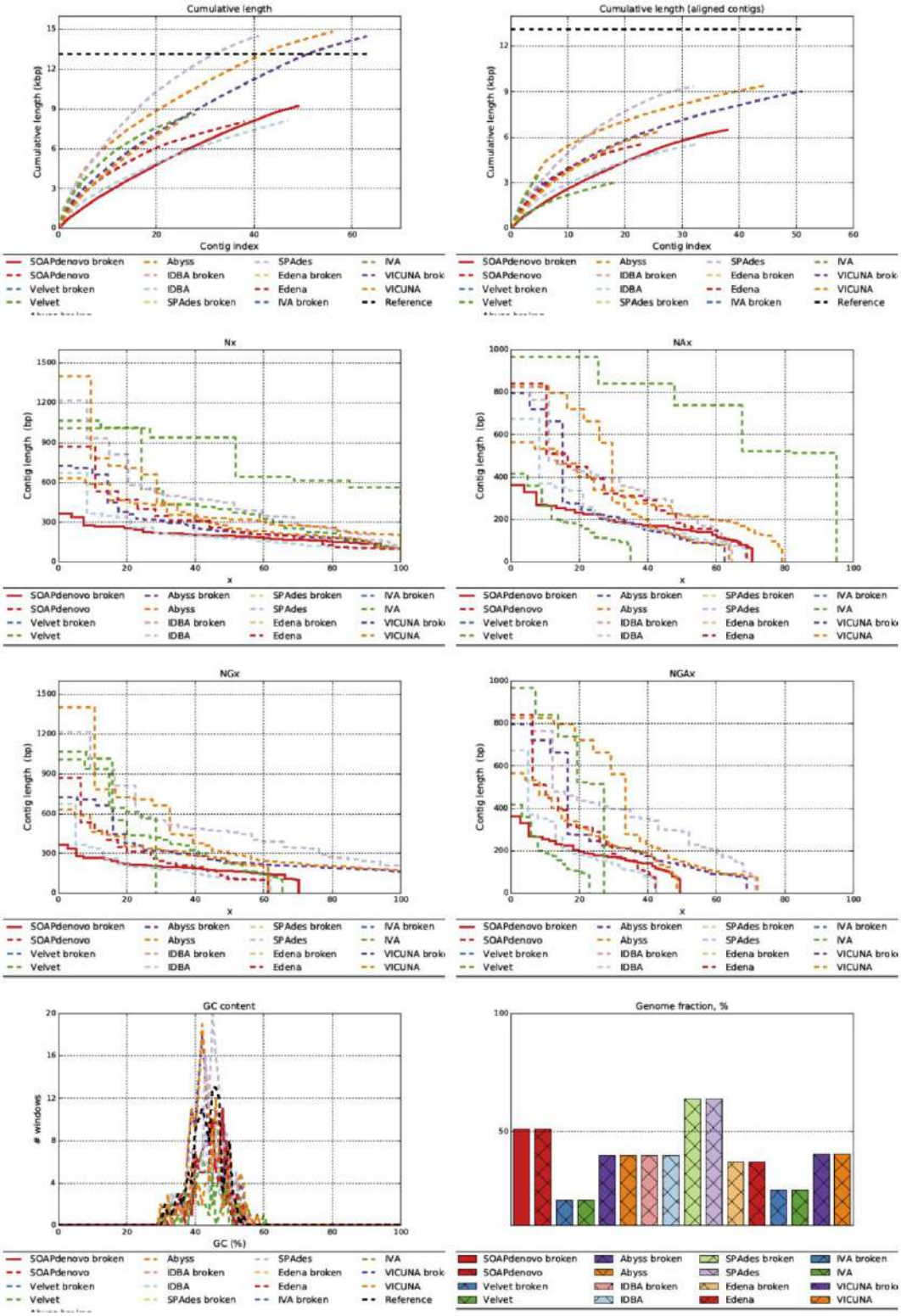


Figure 66. Graphs showing distinct statistics of Influenza_A_GAII assemblies (CY116347)

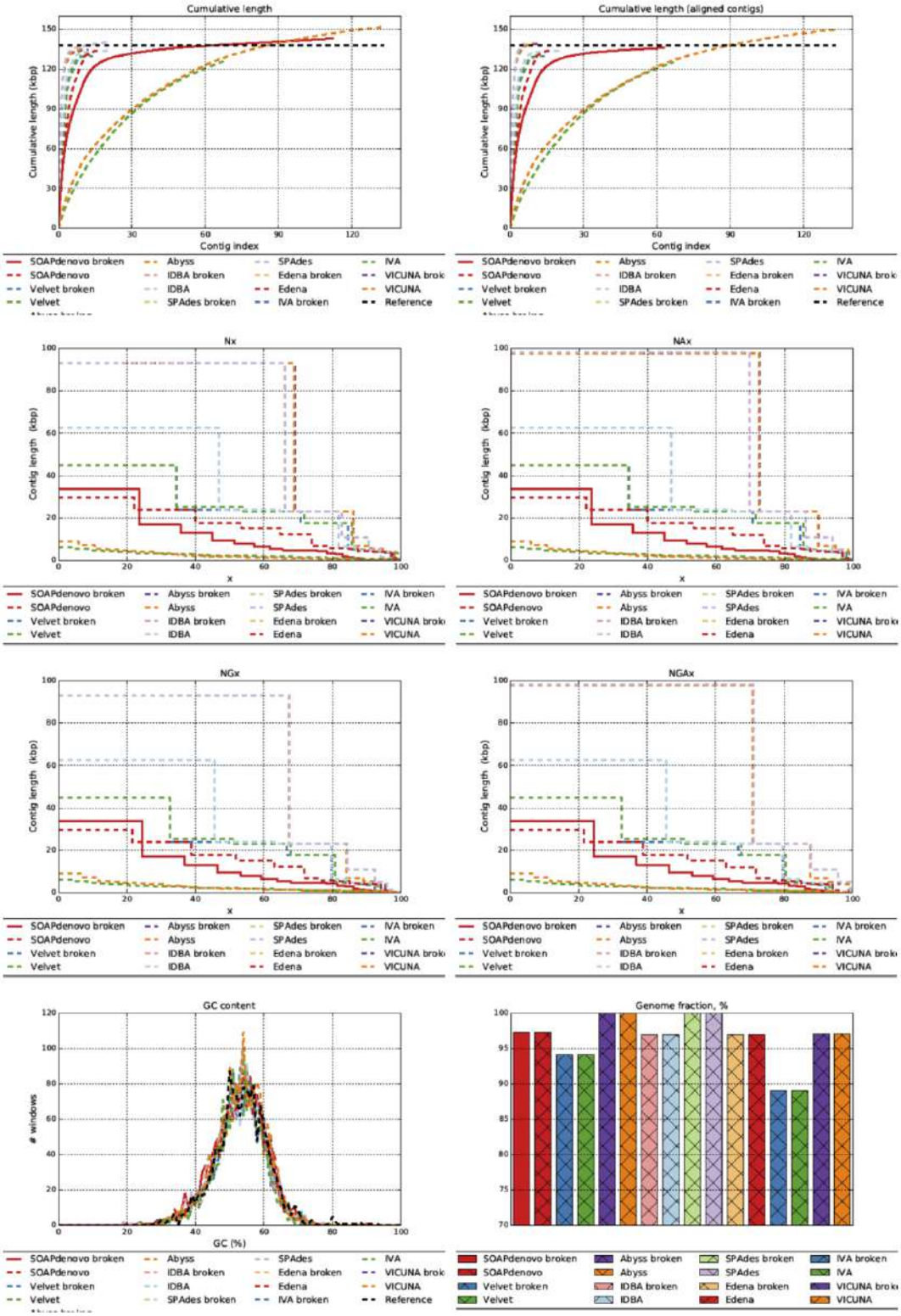


Figure 67. Graphs showing distinct assembly statistics of HHV_8_GAIIx (NC_009333)

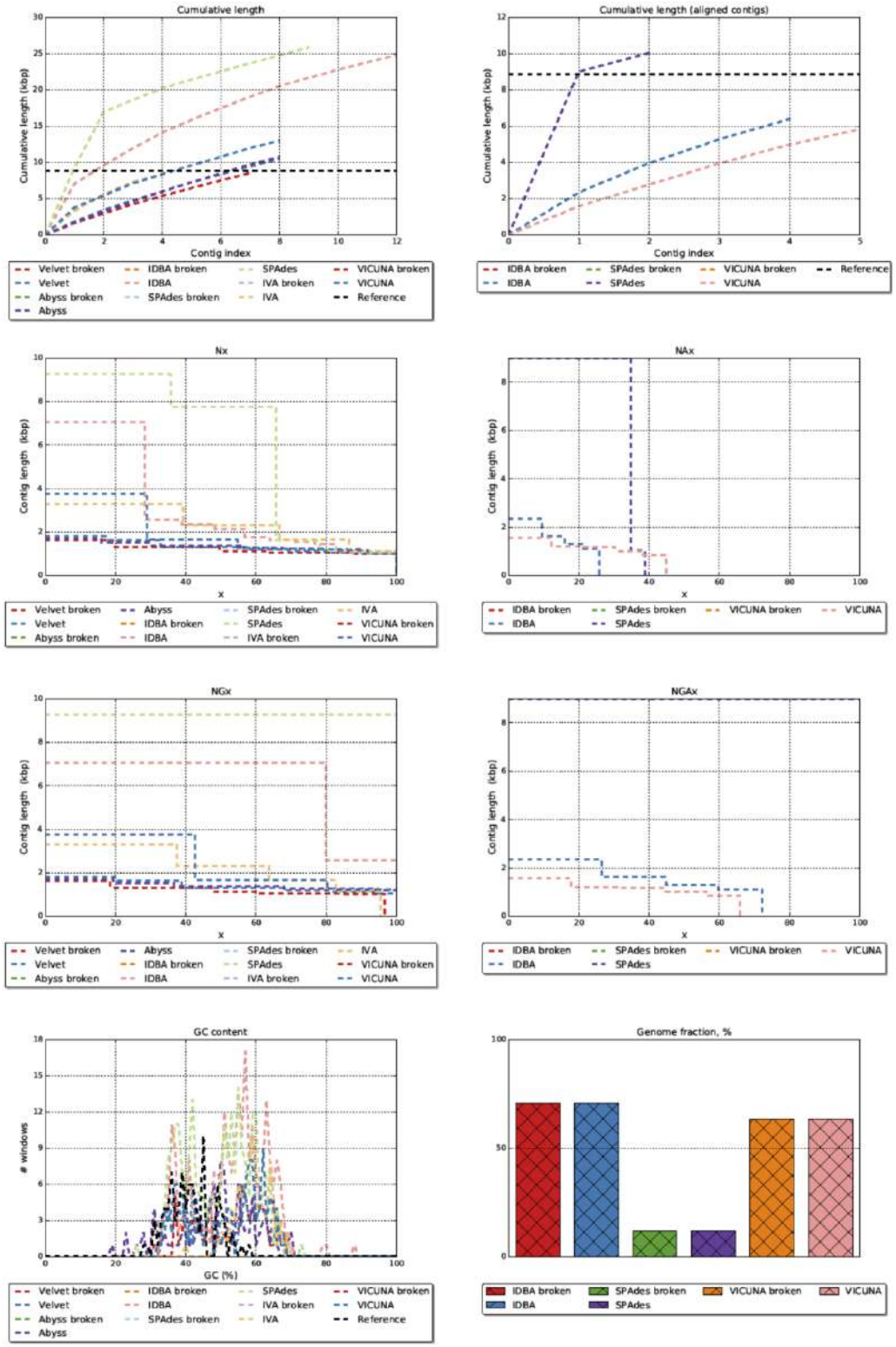


Figure 68. Graphs showing distinct assembly statistics of HIV_1_Hiseq (FJ469707)

Genome assembly of Rhinovirus A (Hiseq) data (SRR499802)

Further, Rhinovirus A Hiseq data of read length 101 bp is assembled utilizing mentioned assemblers. Reference used for assembly assessment is JX074057. We are able to reconstruct Rhinovirus genome upto ~96 % using two assemblers namely IDBA and ABySS. Likewise, other assemblers mainly VICUNA, SPAdes and SOAPdenovo also performed well with the genome fraction percentage of ~88, 83 and 82, respectively (**Table 22**). The least performed tools are IVA and velvet. The largest contig length is obtained using ABySS (6648 bp), IDBA (3906) and SPAdes (2289 bp). Assemblers performance order on Rhinovirus A data is **ABySS**>**IDBA**> VICUNA > SPAdes > SOAPdenovo >**Edena**>**Velvet > IVA**. Performance statistics graphs are depicted in **Figure 69**.

Genome assembly of DENV 3 (Miseq) data (SRR546416)

Genome assembly of DENV 3 data of read length 225 bp is obtained using all the mentioned tools. Overall, all the assemblers performed well on DENV. Reference used for the assessment of DENV 3 assembly is JF920394. We are able to obtain genome fraction of 100% from four assemblers ABySS, SPAdes, VICUNA and IVA followed by Edena (99.4%). The largest contig size of 10682 bp is achieved through ABySS followed by the SPAdes (7721 bp). The least performed assembler is IDBA and velvet with ~97.3% and ~97.6% genome fraction, respectively. The performance of all the assemblers is in the following order **ABySS**>**SPAdes**>VICUNA > IVA > Edena >SOAPdenovo >**Velvet > IDBA.** Quality assessment and statistics are depicted in **Table 22** and **Figure 70**.

Genome assembly of WNV (Miseq) data (SRR546546)

Further, assembly of WNV Miseq data of read length 225 bp is carried out using different assemblers. For the assessment of assembly reference KX547437 is used. Among all, SPAdes, VICUNA and IDBA performed best with the genome fraction percentage of 96.5, 95.3 and 94.6 respectively (**Table 22**). The largest contigs we obtained from the SPAdes of length 5865 bp. The least performed assemblers are IVA and Edena with 20% and 60% genome fraction, respectively. The performance of different assemblers on WNV data is in following order, **SPAdes**>**VICUNA**>**IDBA**> SOAPdenovo > ABySS > Velvet >**Edena**>**IVA**. Performance statistics graphs are depicted in **Figure 71**.

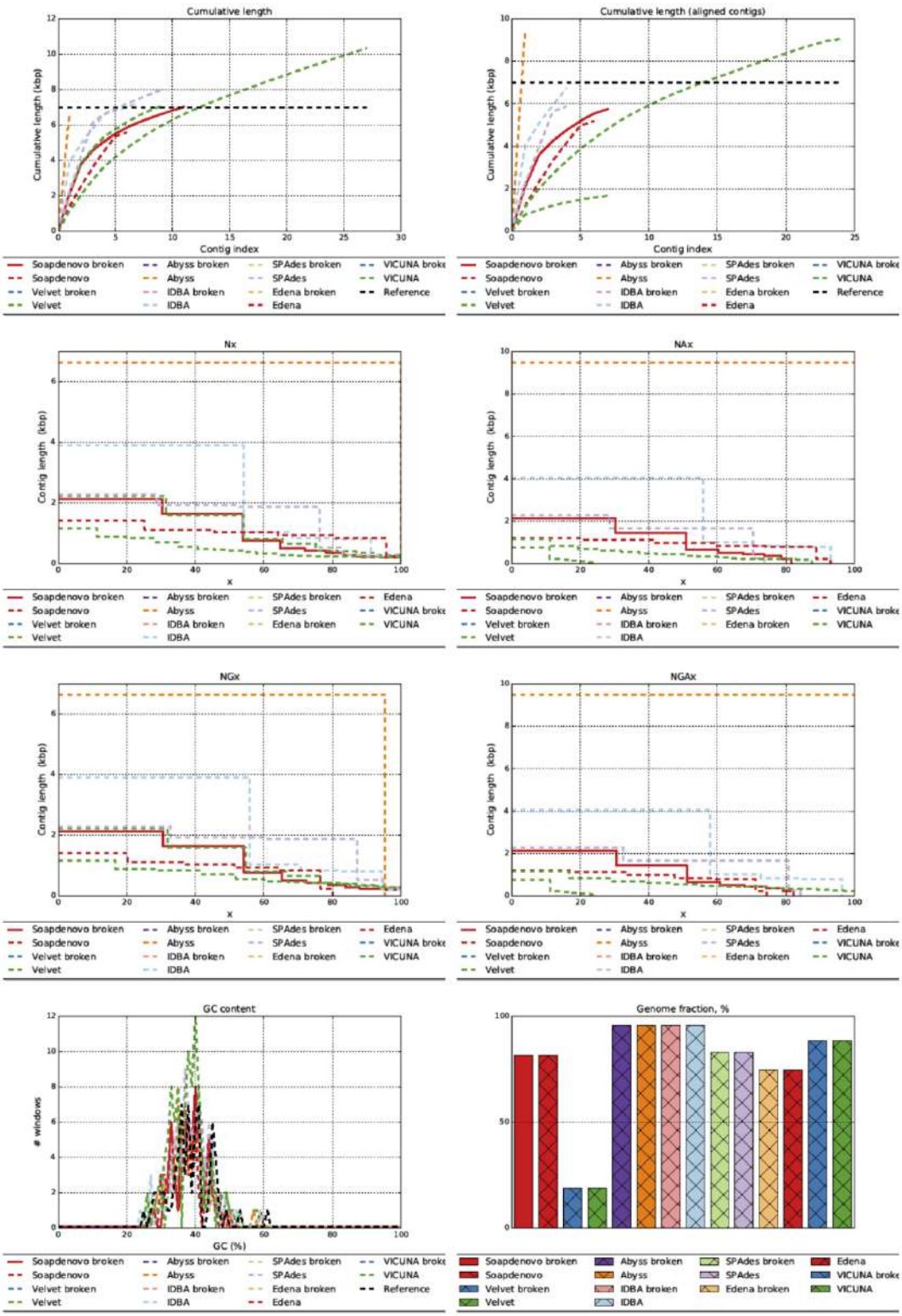


Figure 69. Graphs showing distinct assembly statistics of Rhinovirus_A_Hiseq (JX074057)

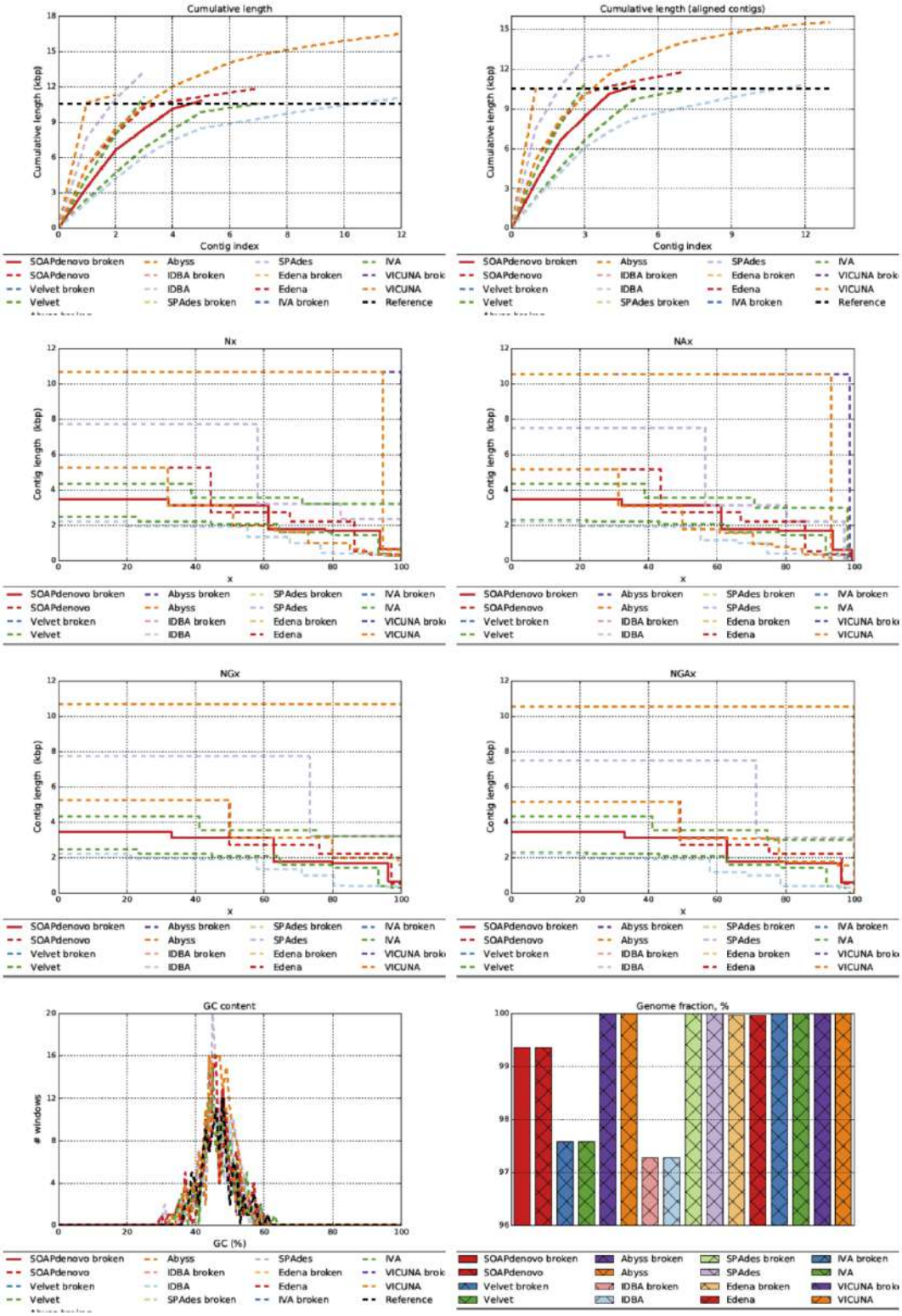


Figure 70. Graphs showing distinct assembly statistics of DENV_3_Miseq (JF920394)

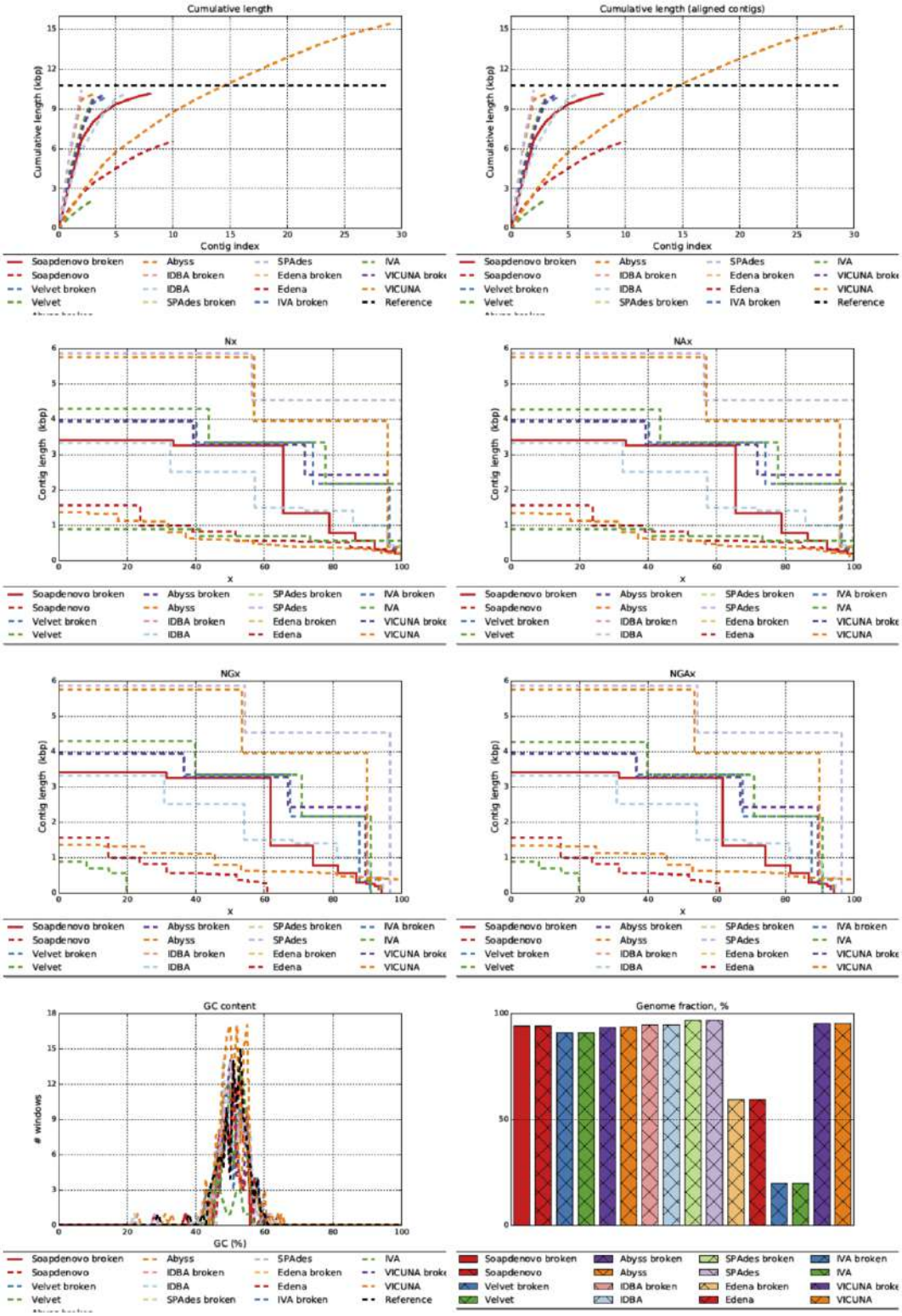


Figure 71. Graphs showing distinct assembly statistics of WNV_Miseq (KX547437)

Genome assembly of HBV (GAIIx) data (DRR001353)

Apart from the paired-end data, the single-end GAIIx HBV data of read length 64 is also analyzed. Reference used for the comparison of assemblies is GQ475322. Among all the assemblers, only IDBA able to reconstruct HBV genome with the 78% genome fraction with largest contig length (4988 bp) followed by SPAdes with 3.4 %. Other assemblers do not perform well. Two assemblers, i.e., IVA and VICUNA do not generate any contig with the defined criteria. Quality assessment and statistics are depicted in **Table 22** and **Figure 72**.

Genome assembly of HPV 16 (NextSeq 500) data (SRR8607785)

Further, we have included the data of HPV 16 from NextSeq platform with the read lengths in the range 149-151. LC511112 is used as reference for the analysis and assessment of distinct assemblies. The best performing assemblers on the data in order to recover HPV 16 genome fraction is VICUNA (80.3%), IDBA (77.8%) and SPAdes (72.9%). The largest contig length is obtained from SPAdes (1499 bp) followed by VICUNA (969 bp) and IDBA (842 bp). The least performing assemblers are IVA, Edena, and velvet. IVA and Edena is not able to generate any contig with defined criteria. The performance order of all the assemblers is **VICUNA**>**IDBA**> **SPAdes**>SOAPdenovo > ABySS >**Velvet > Edena > IVA.** Quality assessment and statistics are depicted in **Table 3**and **Figure 73**.

Genome assembly of SARS-CoV-2 (Miseq) data (SRR11597222)

Moreover, we have also added the latest epidemic SARS-COV-2 Miseq data of read length in the range of 292-301. Reference used for the assessment of different assemblies for SARS-CoV-2 is NC_045512. Among all the assemblers, Vicuna and Edena is not able to generate contigs. The best performing assemblers on the data with the highest genome fraction coverage of 94.6% is obtained through the SPAdes followed by the IDBA with the 92.25% based on the QUAST aligned statistics. Likewise, SOAPdenovo, IVA and Velvet are the least performing assemblers. The largest contig (maximum length) is obtained from the SPAdes of 10482 bp followed by IDBA with maximum contig size of 6375 bp. The performance order of all the assemblers is **SPAdes**>**IDBA> ABySS**> Velvet > IVA>**SOAPdenovo > VICUNA > Edena**. Performance statistics are depicted in **Table 22** and **Figure 74**.

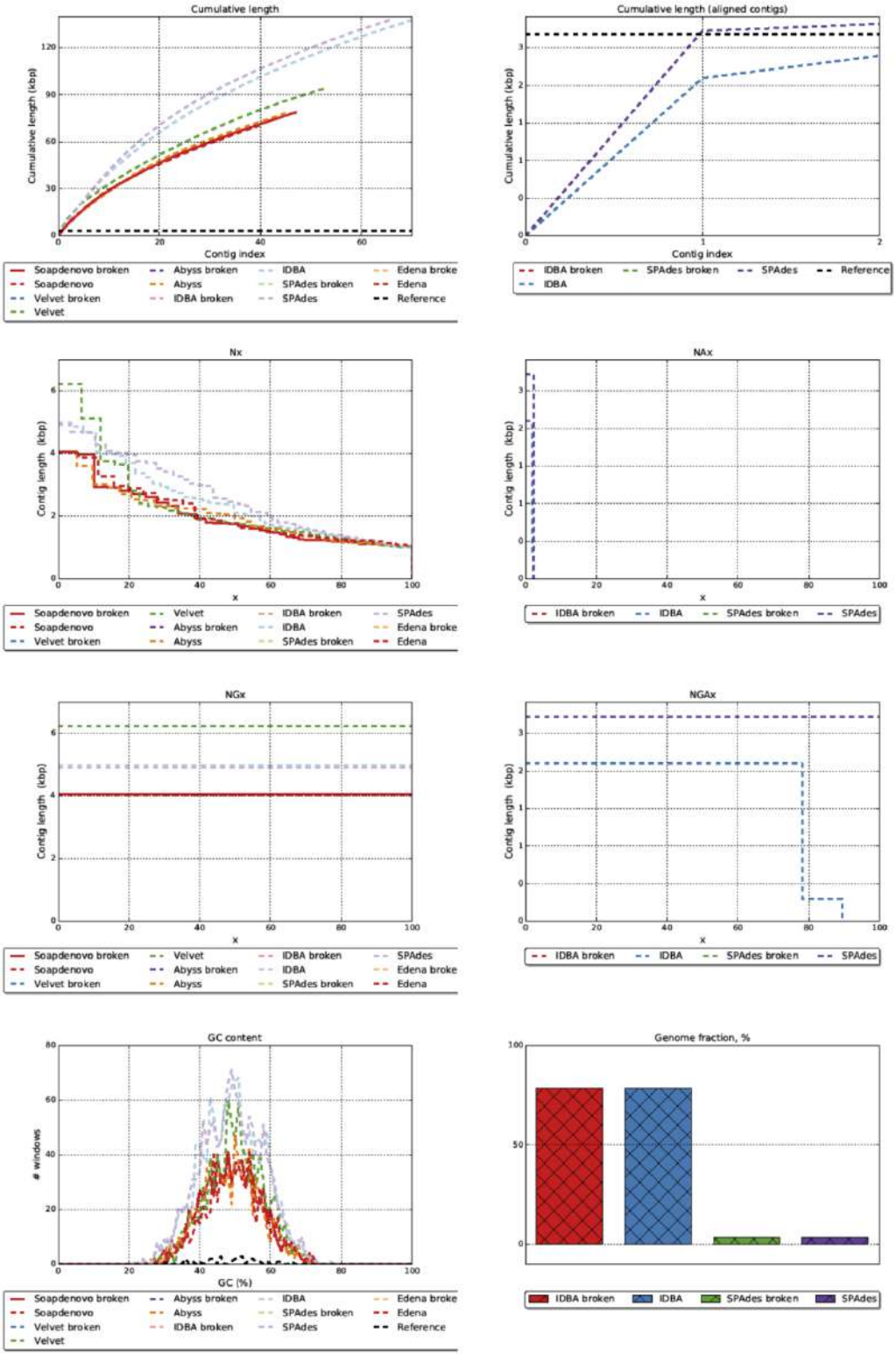


Figure 72. Graphs showing distinct assembly statistics of HBV_GAIIx_single (GQ475322)

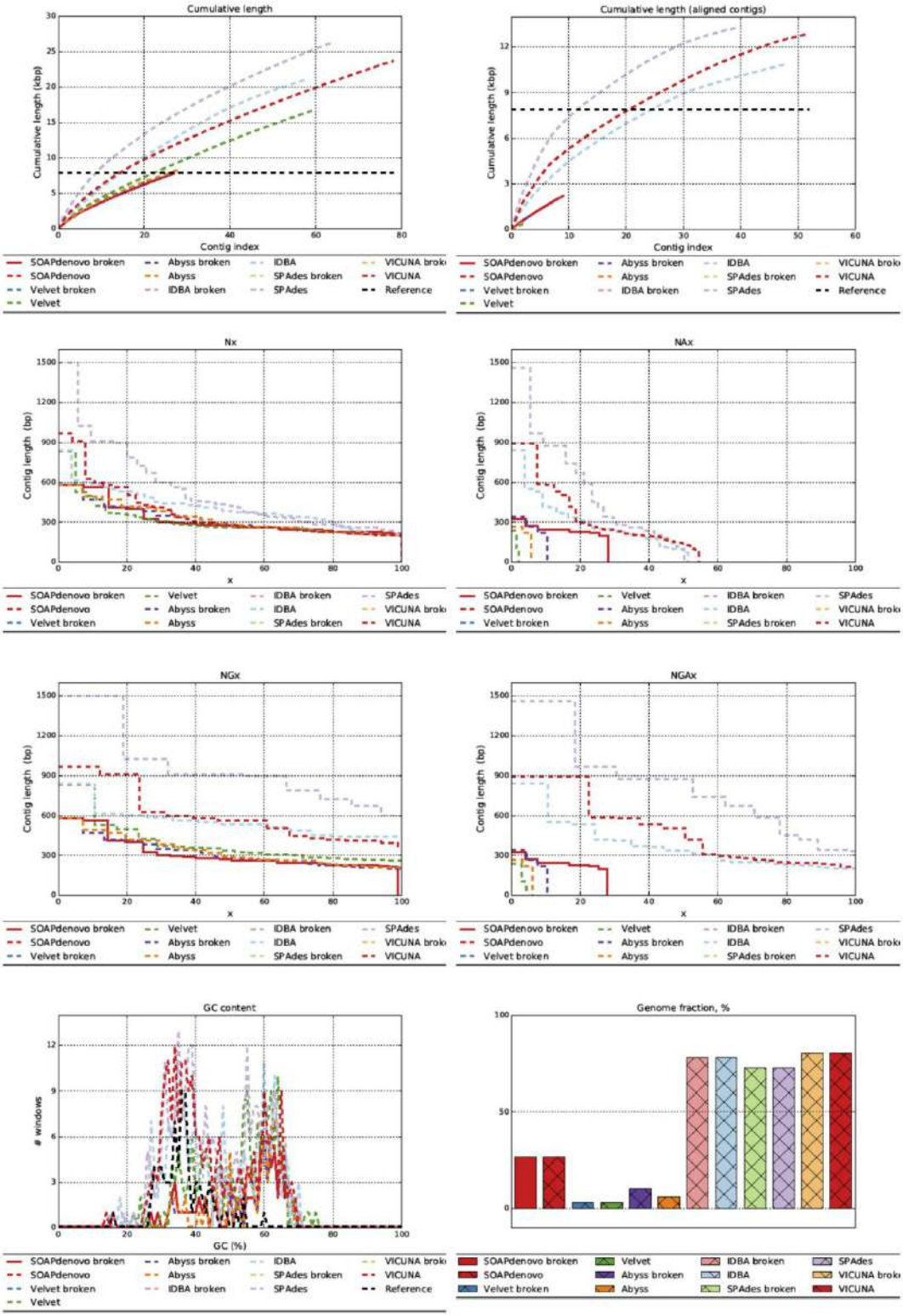


Figure 73. Graphs showing distinct assembly statistics of HPV_16_Nextseq500 (LC511112)

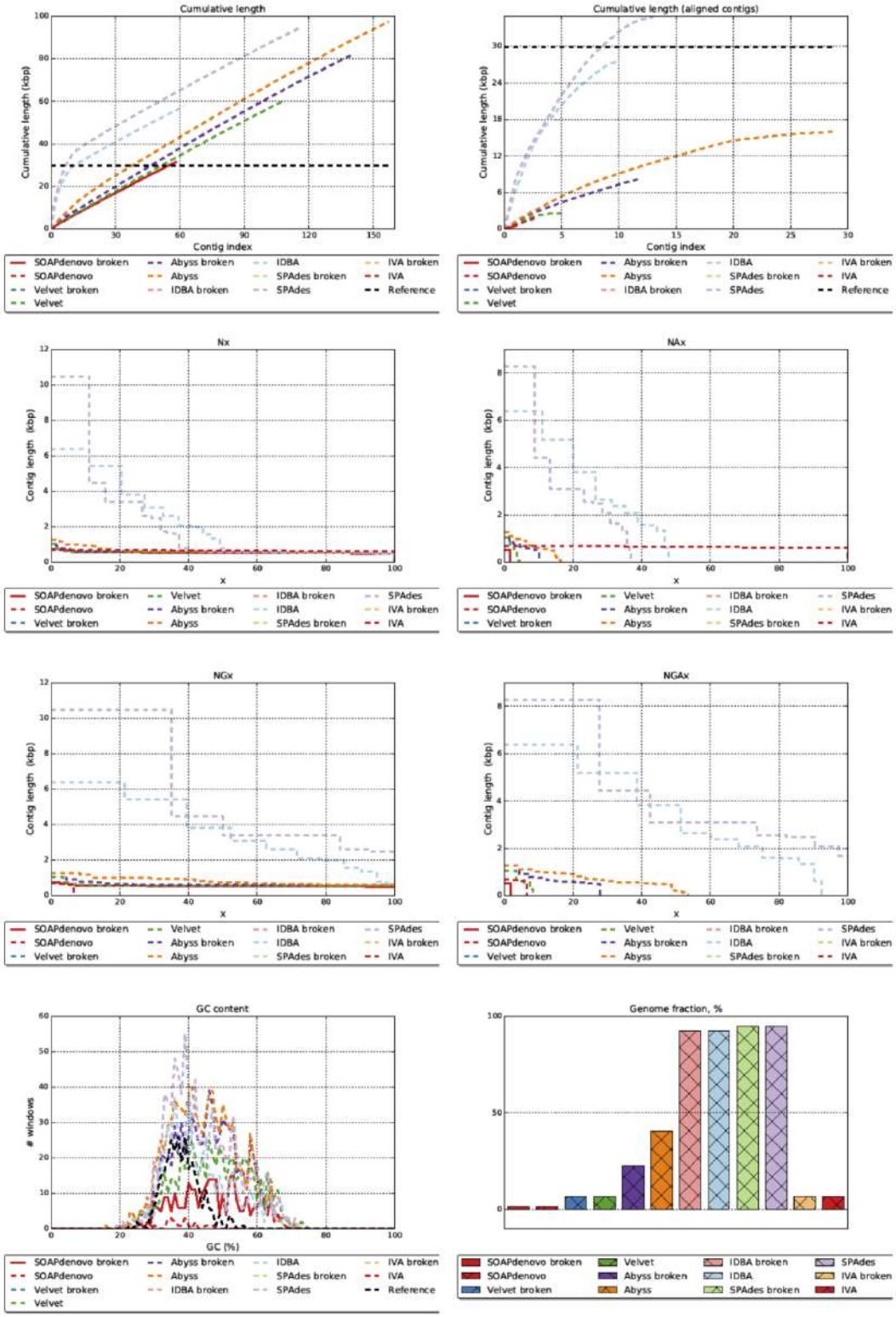


Figure 74. Graphs showing distinct assembly statistics of SARS-2_Miseq (NC_045512)

Conclusion

In the study, we have performed the comparison and assessment of different *de novo* assemblers on the real viral NGS data in order to reconstruct viral genomes. Overall, 8 known assemblers namely, SOAPdenovo, Velvet, ABySS, IDBA, SPAdes, Edena, IVA, VICUNA are included in the study. Likewise, different viral NGS data set from diverse Illumina platforms, i.e., GAII, GAIIx, Hiseq, Miseq, Nextseq 500 is also considered in the work with both paired as well as single-end reads. Based on distinct comparison criteria such as assembly length, assembly aligned length, largest contig length, genome fraction (%) all the assemblers were evaluated and recommendations were made. Here, we have not evaluated the memory and time consumption of the assemblers. Overall, two assemblers, i.e., SPAdes and IDBA performed best in order to recover most of the viral genomes based on the genome fraction percentage. Moreover, existing assemblers performed poor and inconsistent on the retro-viral data. Two viral specific assemblers IVA and VICUNA are also not addressing the underlying aim and not performed best among all, except some exceptions.

Development of bioinformatics tool or pipeline for viral NGS data analysis: Implication in HPV research

Chapter 6. Development of bioinformatics tool or pipeline for viral NGS data analysis: Implication in HPV research

Introduction

Viruses (prokaryotic and eukaryotic) are estimated to be the most abundant biological object on the planet (Paez-Espino et al., 2016). These are causative agent of various deadly diseases and pose risk for human life (Cadwell, 2015; Manso et al., 2017; ME et al., 2013; Paez-Espino et al., 2016; Virgin, 2014; Wolf et al., 2018; Wylie et al., 2012). Further, endogenous viral elements (EVEs) specifically endogenous retroviruses (ERVs) are also very critical, which cover around 5-10% of the Human genome (Campbell et al., 2014; Feschotte and Gilbert, 2012; Holmes, 2011; Horie et al., 2010; Katzourakis and Gifford, 2010). Moreover, prokaryotic viruses (bacteriophages) are known to regulate microbial ecology and play evolutionary role in discrete environments (Clokie et al., 2011; Rohwer, 2003; Wylie et al., 2012). Though, we still have inadequate knowledge regarding the distribution and existence of viruses and it remains to be fully discovered (Shi et al., 2018; Wolf et al., 2018; Woolhouse et al., 2012).

However, next generation sequencing (NGS) and emergence of metagenomics have allowed researchers to explore previously unknown microbial life (viral dark matter) directly from the different environments viz. sediments, soil, seawater, clinical samples, etc. (Foulongne et al., 2012b; Hayes et al., 2017; Manso et al., 2017; Moustafa et al., 2017; Paez-Espino et al., 2016; Reyes et al., 2010; Shi et al., 2018; Wylie et al., 2014). Moreover, virome studies have drastically expanded the viral genomic sequences (Delwart, 2007; Edwards and Rohwer, 2005; Kristensen et al., 2010; Mokili et al., 2012).

To fulfill the unprecedented need to explore virome, assorted resources and computational methods have been developed utilizing combinations of different tools (Fancello et al., 2012; Nooij et al., 2018; Orton et al., 2016). Like, Roux et. al. developed a webserver Metavir and Metavir2 for analysis of viral metagenomes (Roux et al., 2011; Roux et al., 2014). Wommack et al. established a resource VIROME for viral metagenome sequence exploration (Wommack et al., 2012). Ho and Tzanetakis et. al. developed VirFind for virus detection (Ho and Tzanetakis, 2014). Rampelli et al.

constructed ViromeScan tool for viral (eukaryotic viruses) community profiling from metagenomic reads (Rampelli et al., 2016). Similarly, Li et al. developed computational pipeline VIP for virus identification and discovery (Li et al., 2016c). Likewise, Zhao et al. developed VirusSeeker for virus discovery (Zhao et al., 2017). Lin et al. provide web-based pipeline Vipie for viral characterization from NGS samples (Lin et al., 2017). Tithi et al. describe FastViromeExplorer, a pipeline for identification of viruses and phages (Tithi et al., 2018). Maarala et al. provide ViraPipe, a scalable and cluster-based pipeline for viral metagenomic analysis (Maarala et al., 2018). Garretto et al. developed virMine for automatic detection of viral sequences from metagenomic samples (Garretto et al., 2019). Moreover, different algorithms were also developed specifically to detect (pro)phages such as PHASTER (Arndt et al., 2016), MetaPhinder (Jurtz et al., 2016), etc. Simultaneously, distinct tools like SURPI (Naccache et al., 2014), Clinical PathoScope (Byrd et al., 2014), PathoScope (Hong et al., 2014) were also developed to detect pathogens in clinical samples. Furthermore, Greninger et al. developed MetaPORE to identify viral pathogens in clinical samples from nanopore sequencing (Greninger et al., 2015).

However, there are different challenges exist in the analysis of viral metagenomic data (Lambert et al., 2018; Rose et al., 2016). Like, some of the algorithms only available standalone, either works with environmental or clinical data, missing quality control step, works with single sequencing platform, need computing expertise etc. Here, we are demonstrating an integrated ready-to-use standalone and online computational pipeline, VIRpipe, for the identification and analysis of viral abundance from metagenomic raw data from Illumina paired-end sequencing as well as nanopore technology data.

Materials and Method

Installation and configuration (Dependencies)

Different external programs and tools (dependencies) were utilized, and integrated (**Table 23**) to develop VIRpipe.

Table 23. All the dependencies (OS, programming languages, program, tools) employed in the VIRpipe

OS/Language/Program/Tools	***Application***	***References***
Linux (Ubuntu)	Operating system	https://www.linux.org/, https://ubuntu.com/
Bash	Shell scripting	https://www.gnu.org/software/bash/
Perl (GD, GD::Graph)	Scripting language and visualization	https://www.perl.org/
R	Scripting, statistical computing and visualization	https://www.r-project.org/
PHP	Scripting and web development	https://www.php.net/
HTML (CSS/JS)	Web development	https://www.w3.org/
LAMP	Webserver hosting	https://ubuntu.com/server/docs/lamp-applications
R Markdown	Report and presentation	https://rmarkdown.rstudio.com/
NGSQC-Toolkit	Quality Control	(Patel and Jain, 2012)
Poretools	Quality filtering and read extraction	(Loman and Quinlan, 2014)
BWA-mem	Sequence mapping	(Li and Durbin, 2010)
BLAST	Sequence mapping	(Camacho et al., 2009)
Usearch v11/Ublast (32 bit)	Sequence mapping	(Edgar, 2010)
SAMtools	File processing and conversions	(Li et al., 2009)
Krona tool	Visualization	(Ondov et al., 2011)

VIRpipe and web-portal

An integrative ready-to-use virome profiling pipeline utilizing bash scripting, Perl/bioperl, R/Bioconductor, is developed. Various programs and tools, i.e., Samtools (Li et al., 2009), Poretools (Loman and Quinlan, 2014), NGSQC-toolkit (Patel and Jain, 2012), BWA (Li and Durbin, 2010), Blastn (Camacho et al., 2009), Ublast (Edgar, 2010) and Krona (Ondov et al., 2011) is deployed (**Table 23**). Additionally, web service and platform is also developed. It is implemented using the Linux-Apache-MySQL-PHP (LAMP) open source solution bundle with combination of programming and scripting languages, i.e., HTML, PHP, Perl and R. It will be freely available for wide scientific community engaged in viral informatics and viromics.

Results and Discussion

Sequence resources and indexing

Different sequence (nucleotide (nt) and protein (pro)) resources is created for VIRpipe (**Table 24**). Further, indexing of these sequence resources (**Table 24**) is performed to utilize in VIRpipe mapping steps. Different indexed resources are as follows:

1. VIRdb: This is the compendium of viral reference genomes that includes 9594 viral genome sequences at the time of development from NCBI

2. VIRproDB: This is the repository of RefSeq viral protein sequences consist of 477628 sequences from NCBI

3. HVPCdb: It is generated from the Human Virome Protein Cluster database sequences (Elbehery et al., 2018). This provide representative ORFs from the human virome data of six body sites

4. EVEntDB/EVEproDB: It is generated from the Genome-based Endogenous Viral Element Database (gEVE v1.1) (Nakagawa and Takahashi, 2016). This includes sequences of endogenous viral elements (EVEs) with endogenous retroviruses (ERVs) from Human.

Table 24. Sequence resources (nucleotide and protein) and indexing status

Indexed resource	*Sequences*	*BWA*	*BLAST*	*Ublast*
VIRdb	9594	✓	✓	
VIRproDB	477628			✓
HVPCdb	390917			✓
EVEntDB/EVEproDB	33966			✓

nt: Nucleotide, pro: Protein

Raw data processing (Quality control and read extraction)

Raw metagenomic sequencing data from both the platforms (Illumina and Nanopore) were quality filtered and extracted into fasta/fastq file(s). For Illumina data, NGSQC Toolkit is employed for the removal of low-quality reads (default PHRED score < 30), sequencing errors and adapter (inbuilt or user provided adapters) contamination. In case of Nanopore (MinION), all the high quality 2D reads from raw sequencing data (FAST5) will be extracted in to fasta/fastq files utilizing Poretools.

VIRpipe: Virome identification and profiling

The complete pipeline is divided into two separate modules, i.e., clinical module (CM) and environmental module (EM). In the clinical module, metagenomic-sequencing data from the Human host can be analyzed for viral identification and exploration. Likewise, environmental module will be used for the analysis of metagenomic data from the different environments, i.e., sediments, water, soil etc. Simultaneously, both the sequencing platforms viz. Illumina and Nanopore are supported. The complete outline and structure of VIRpipe is depicted in **Figure 75**.

For rapid viral identification and profiling, we have opted for sequential search space reduction approach to maintain time and sensitivity. First, VIRpipe perform fast and less sensitive (relaxed) mapping (M-1) of quality filtered reads utilizing BWA-mem algorithm and VIRdb BWA-indexed resource. Default mapping criteria for Illumina data and predefined optimal (-ont2d) settings were utilized for nanopore sequencing data. Further, it processes the sequence alignment map (SAM)/binary alignment map (BAM) files to retrieve all the mapped reads (named as probable viral reads) utilizing SAMtools. All the probable viral reads (reduced) from M-1, were again subjected to the sensitive mapping (M-2) using Blastn program and VIRdb BLAST-indexed resource. For this, strict criteria are defined to avoid the false positive abundance. In case of

Illumina data, 90% identity and 70% coverage constrain with one max target (best hit) was used. Likewise, 70% identity and coverage criteria with one max target for Nanopore data was set. Then, Unmapped reads from M-2 (UnM-2) were further subjected for mapping (M-3) against protein resources employing Usearch/Ublast (32 bit) algorithm. In the clinical module (CM), both the resources, i.e., VIRproDB and HVPCdb will be used. Similarly, in environmental module (EM) only VIRproDB will be utilized. Finally, all the mapped reads from M-2 and M-3 will be designated for complete virome distribution.

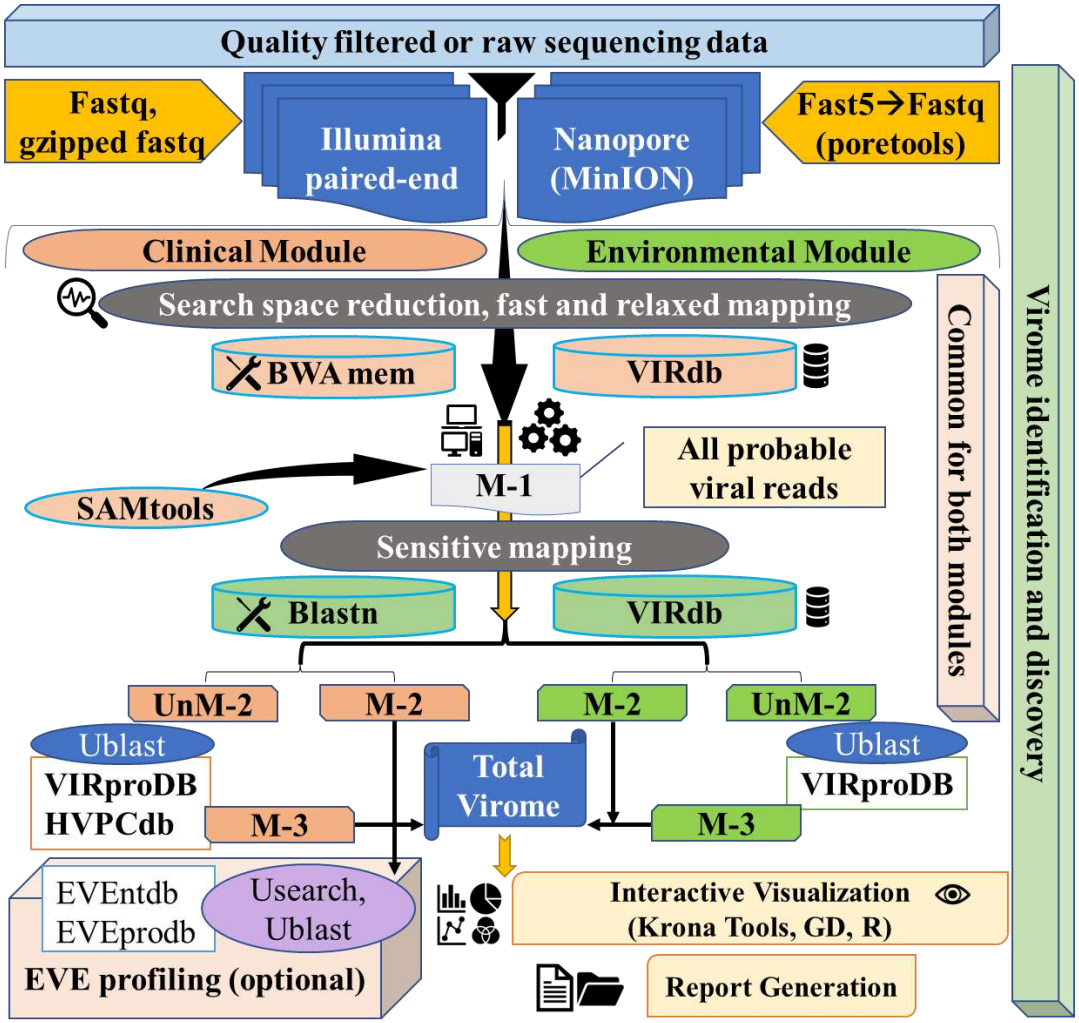


Figure 75. Complete outline of VIRpipe

Further, we also provide EVE profiling module to explore the distribution of potential endogenous viral elements (EVEs) with endogenous retroviruses (ERVs) from identified virome (mapped reads from M2+M3) in Human clinical samples. At last, interactive visualization and report will be generated of the analysis.

VIRpipe-Online web-portal

In addition to standalone version, an online web-portal is also developed to provide ready-to-use option for VIRpipe. It is freely available at https://bioinfo.imtech.res.in/manojk/virpipe/. The standalone version and all the pre-built index files (databases) were also provided online to download. To use VIRpipe-online users has to perform first step of quality control locally. Commands to utilize for this is also provided in the manual available online. The third-party links to dependencies used to develop VIRpipe is also provided at the server.

To use VIRpipe-online, Users has to opt for the appropriate options according to the underlying samples (clinical or environmental) and sequencing technology (Illumina or Nanopore). Further, user has to upload quality filtered data (Fastq) files to the server and run the VIRpipe through submit button to explore and identify virome. Users can also use EVE profiling module (optional) and perform EVE analysis. Complete analysis result and report will be available at the respective link for three days to visualize or download.

VIRpipe approach:

Input: Raw sequencing data (fastq) file(s)
Sample source (Clinical or Environmental)
Sequencing platform
Result: Virome identification and distribution, Endogenous viral elements (optional)
if M=CM then # Clinical module
if P=I then # Illumina data
Quality Control and preprocessing ~NGSQC Toolkit # Raw reads (Fastq files)
Fast and relaxed mapping (M-1) ~BWA-MEM (*VIRdb*)
1st-tier space reduction
Probable viral reads extraction ~Samtools
Strict and sensitive mapping (M-2) ~blastn (*VIRdb*)
2nd-tier space reduction
Mapped and Unmapped reads extraction ~Samtools
Unmapped reads mapping to protein resources (M-3) ~Ublast (*VIRproDB, HVPCdb*)
else # Nanopore data
2D high quality reads extraction ~poretools # Raw reads (FAST5)
Fast and relaxed mapping (M-1) ~BWA-MEM (-ont2d, *VIRdb*)
Probable viral reads extraction ~Samtools
Strict and sensitive mapping (M-2) ~blastn (*VIRdb*)
Mapped and Unmapped reads extraction ~Samtools
Unmapped reads mapping to protein resources (M-3) ~Ublast (*VIRproDB, HVPCdb*)
end
if EVE=1 then #EVE profiling optional
Mapping M2+M3 mapped reads ~Ublast (EVEntDB)
Mapping Unmapped reads ~Ublast (EVEproDB)
end
report generation
else #Environmental module (M=EM)
if P=I then # Illumina data
Quality Control and preprocessing ~NGSQC Toolkit # Raw reads (Fastq files)
Fast and relaxed mapping (M-1) ~BWA-MEM (*VIRdb*)
1st-tier space reduction
Probable viral reads extraction ~Samtools
Strict and sensitive mapping (M-2) ~blastn (*VIRdb*)
2nd-tier space reduction
Mapped and Unmapped reads extraction ~Samtools
Unmapped reads mapping to protein resources (M-3) ~Ublast (*VIRproDB*)
else # Nanopore data
2D high quality reads extraction ~poretools # Raw reads (FAST5)
Fast and relaxed mapping (M-1) ~BWA-MEM (-ont2d, *VIRdb*)
Probable viral reads extraction ~Samtools
Strict and sensitive mapping (M-2) ~blastn (*VIRdb*)
Mapped and Unmapped reads extraction ~Samtools
Unmapped reads mapping to protein resources (M-3) ~Ublast (*VIRproDB*)
end
report generation
end

Case study: Virome Profiling from anogenital warts

Further, we have implemented VIRpipe to investigate viral metagenomic data from anogenital warts and rapidly identify viral abundance. Sequencing data is retrieved from the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/home). For this, two viral metagenomic data were retrieved. The data belong to the study accession number: PRJNA517793. Both the data is from the Illumina MiSeq platform with paired-end reads (250 bp) generated from the pooled specimens each with 10 warts samples. Detail about metagenomic data is provided in **Table 25**.

Table 25. Viral metagenomic data from anogenital warts

No.	Run accession	Experiment accession	Read Counts
1	SRR8509862	SRX5313520	122008
2	SRR8509868	SRX5313526	48561

Virome from SRR8509862 data

In rapid screening, 13875 sequences (reads) belong to 44 viruses were identified in the anogenital warts metagenomic data SRR8509862 (**Table 26**). Among these, abundance of sequences belonging to 35 viruses is very low, which will need further attention to characterize. However, it is also very critical to identify less abundant viruses in the sample. Thirteen viruses were found in abundance with 3 HPVs (6b, 53, and 85) (**Figure 76**). We have mainly identified different HPV types from the metagenomic data as HPVs mainly play role in the genital warts. Importantly, 11 papillomaviruses were identified, among which 10 are human papillomaviruses (type 6b, 53, 85, 54, 90, 103, 16, 129, 92, and 7). Krona plot depicting *Alphapapillomaviruses* in anogenital warts virome (**Figure 77**). The most abundant virus demarcated is HPV 6b (45%) that is low-risk HPV primarily known to cause warts. Also, 23 phages were recognized, which are largely characterized as Staphylococcus phages (15) (**Table 26**).

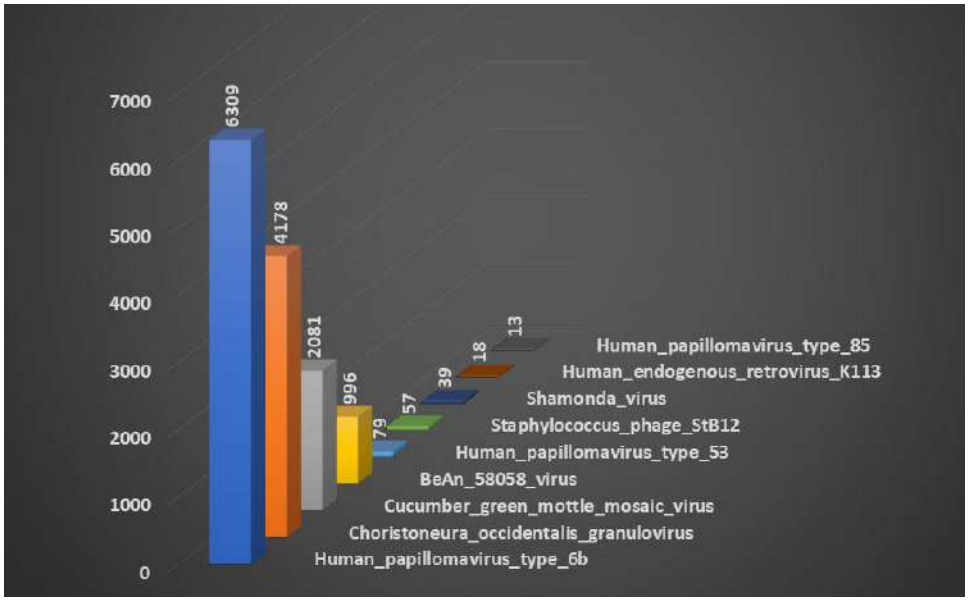


Figure 76. Thirteen most abundant viruses with number of designated reads in SRR8509862 data

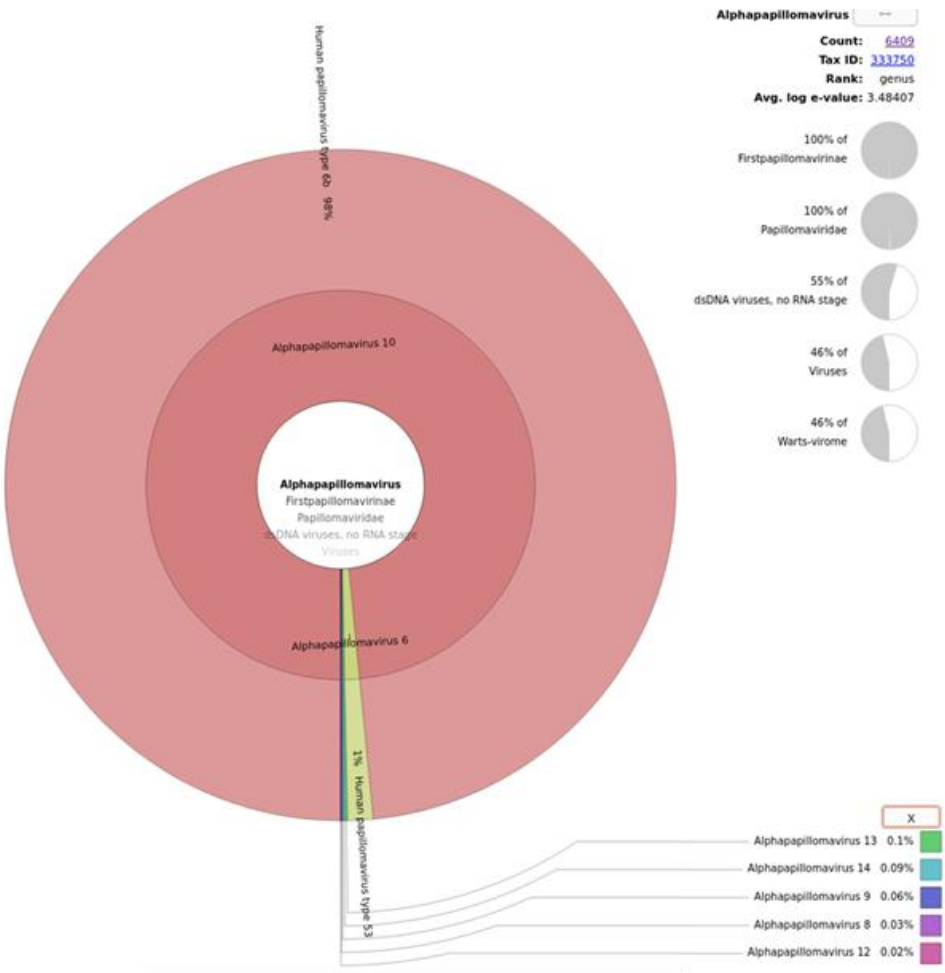


Figure 77. Krona plot depicting *Alphapapillomavirus* component in warts virome (SRR8509862)

Table 26. List of 44 viruses identified in metagenomic (SRR8509862) data

Viruses	***Read Count (Abundance)***
Human_papillomavirus_type_6b	6309
Choristoneura_occidentalis_granulovirus	4178
Cucumber_green_mottle_mosaic_virus	2081
BeAn_58058_virus	996
Human_papillomavirus_type_53	79
Staphylococcus_phage_StB12	57
Shamonda_virus	39
Human_endogenous_retrovirus_K113	18
Human_papillomavirus_type_85	13
Staphylococcus_phage_StB20-like	9
Human_papillomavirus_54	8
Human_papillomavirus_type_90	6
Human_papillomavirus_type_103	6
Staphylococcus_phage_StB20	5
Tokyovirus_A1_nearly	4
Staphylococcus_phage_vB_SepiS-phiIPLA7	4
Staphylococcus_phage_vB_SepiS-phiIPLA5	4
Staphylococcus_aureus_phage_phiNM2	4
Lactobacillus_phage_AQ113	4
Human_papillomavirus_type_16	4
Human_papillomavirus_type_129	4
Staphylococcus_phage_PH15	3
Gordonia_phage_Yeezy	3
Torque_teno_virus_8	2
Torque_teno_virus_1	2
Streptococcus_phage_SMP	2
Staphylococcus_prophage_phiN315	2
Staphylococcus_phage_X2	2
Staphylococcus_phage_tp310-3	2
Staphylococcus_phage_phiRS7	2
Staphylococcus_phage_IME-SA4	2
Staphylococcus_phage_CNPx	2
Staphylococcus_phage_187	2
Salmonella_phage_SJ46	2
Pepper_mild_mottle_virus	2
Human_papillomavirus_type_92	2
Human_papillomavirus_type_7	2
Escherichia_phage_TL-2011b	2
Enterobacteria_phage_13a	2
Taterapox_virus	1
Staphylococcus_phage_CNPH82	1
Rhodococcus_phage_ReqiPoco6	1
Rhodococcus_phage_ReqiPepy6	1
Papio_hamadryas_papillomavirus_type_1	1

Virome from SRR8509868 data

Similarly, metagenomic data, i.e., SRR8509868 is also analyzed for virome identification. In this, 6974 sequences were delineated into 15 viruses including 9 with very low abundance (**Table 27**). This is again dominated by papillomaviruses with 7 HPVs (6b, 32, 7, 61, 53, 16, and 54). In line, most dominating virus among all is HPV 6b, approximately 71% (4964) of total designated sequences. Abundance of *Alphapapillomaviruses* is shown in **Figure 78**. However, we have not identified any phage in the data.

Table 27. List of 15 viruses identified in metagenomic (SRR8509868) data

Viruses	***Read Count (Abundance)***
Human_papillomavirus_type_6b	4964
Choristoneura_occidentalis_granulovirus	1594
BeAn_58058_virus	360
Human_papillomavirus_type_32	12
Shamonda_virus	8
Human_endogenous_retrovirus_K113	8
Human_papillomavirus_type_7	6
Chicken_picornavirus_5	6
Human_papillomavirus_-_61	4
Rhesus_monkey_papillomavirus	2
Melegrivirus_A	2
Human_papillomavirus_type_53	2
Human_papillomavirus_type_16	2
Human_papillomavirus_54	2
Ferret_papillomavirus	2

Furthermore, various studies also suggest that almost all types of virus including retroviruses can be endogenized in the host genome including Human (Feschotte and Gilbert, 2012; Horie et al., 2010; Katzourakis and Gifford, 2010). Identification of EVEs can also be useful in understanding evolution, and finding reservoirs (Feschotte and Gilbert, 2012; Holmes, 2011; Ueda et al., 2020). Additionally, EVEs can also play role in genome structure and distinct physiological functions including genomic instability, recombination, etc. along with diseases (Campbell et al., 2014; Jern and Coffin, 2008; Küry et al., 2018; Xue et al., 2020). However, various viral identification methods use host subtraction approach, which takes large amount of memory and computational time. Moreover, host (Human) sequence removal step may also create blind spot, i.e., lead towards loss of endogenous viral elements (sequences) and endogenous retroviruses of interest (Lambert et al., 2018). Hence, VIRpipe EVE-module will be unique and useful to explore endogenous viral elements.

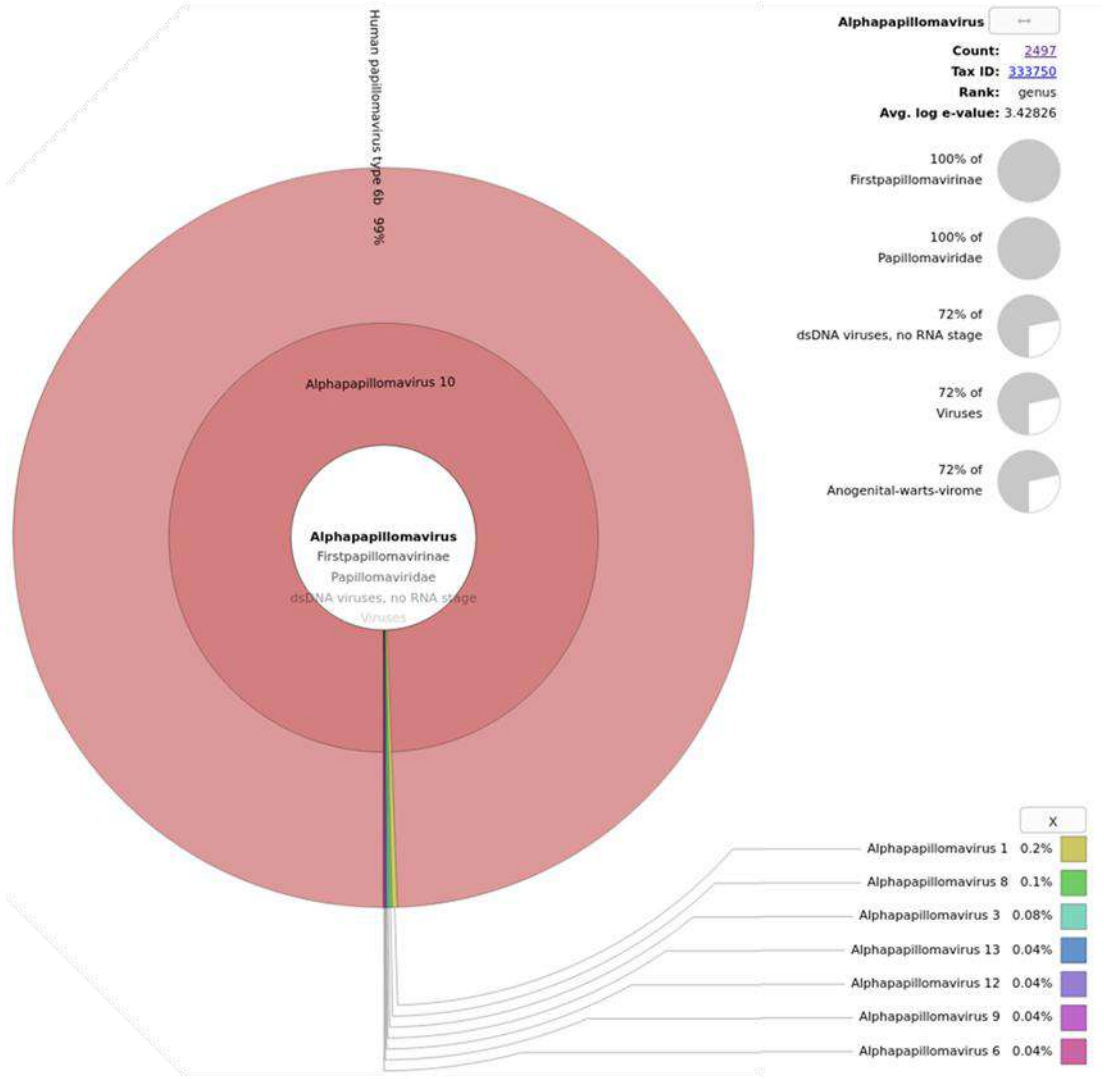


Figure 78. Krona plot showing *Alphapapillomavirus* abundance in warts virome (SRR8509868)

Conclusion

Viruses are most abundant biological entity on earth. Though, the large portion is still unknown and undiscovered. However, viral metagenomics provide promising tool to explore viruses from distinct ecosystem. We have developed, VIRpipe; an integrated ready-to-use pipeline for rapid virus identification from the clinical and environmental metagenomic samples using raw data from both Illumina paired-end as well as nanopore sequencing data. Additionally, VIRpipe can also use to identify human endogenous retroviruses in clinical samples.

Summary, Future Implications and Directions

Chapter 7. Summary, Future Implications and Directions

Summary

Human papillomaviruses (HPVs) are the double-stranded DNA (dsDNA) oncogenic viruses belong to the *Papillomaviridae* family. They are known to cause numerous carcinomas mainly cervical, head and neck, vulvar, penile, etc. Oncogenesis includes series of steps, mainly HPV infection progress from persistent infection to pre-cancer and invasive carcinoma along with series of possible events in process. Regardless of substantial achievements, there is prerequisite for the establishment of effectual biomarkers to distinguish disease progressions. Thus, we have developed a comprehensive resource "HPVbase" of potential biomarkers for HPV mediated diseases. This includes viral integration and breakpoint events, HPVs methylation patterns and HPV mediated aberrant expression of distinct host microRNAs (miRNAs). It comprises of 1257 integration events from distinct HPV types mainly 16 (954), 18 (216), 33 (33) and 45 (33) related with different histological circumstances. HPV integrant browser is also constructed. Correspondingly, it also contains 719 quantitative HPV DNA methylation entries pertaining to 5 HPV genotypes namely HPV 16 (495), HPV 18 (113), HPV45 (66), HPV 31 (34) and HPV 33 (11). Furthermore, aberrant expression profile of 341 miRNAs from diverse carcinoma along with their target genes were curated and compiled that can be useful for miRNA-based therapeutics. For easy data access and retrieval, a user-friendly web interface has been developed. We anticipate that HPVbase would assist the scientific community engaged in HPV research. Complete resource is freely available at http://crdd.osdd.net/servers/hpvbase**.**

HPVs are also divided into the high-risk (HR), and low-risk (LR) based on their disease-causing competence. Persistent infection of high-risk HPVs is primarily associated with the carcinogenicity. Irrespective to molecular advancements in screening, vaccination, and prevention strategies, cervical carcinoma remains to be fourth most common cancer among women. Therefore, there is still requirement for the identification of potential targets and drugs effective to inhibit HPV infection and cancer progression. For this, multi-omics analysis was performed on the compendium of 1887 HPV infection-associated and HPV integration driven disrupted genes. Our illustrative approach revealed potential key therapeutic targets, hallmark molecular

functions and enriched pathways, transcription factors, microRNAs, genomic alterations, and potential drug candidates to explicate HPV pathogenesis. Importantly, significant enrichment of hallmarks and pathways in cancer, viral carcinogenicity, Human papillomavirus infection, G2M checkpoint, E2F-Targets, Apoptosis, EGFR tyrosine kinase inhibitor resistance, PI3K_AKT_MTOR signaling, IL6_JAK_STAT3 signaling, ErbB signaling, Epithelial mesenchymal transition, P53 pathway, DNA repair, Spermatogenesis, and NOTCH signaling is identified. Genomic alteration profiling further substantiated our findings. Among identified key targets, TP53, NOTCH1, PIK3CA, EP300, CREBBP, EGFR, ERBB2, PTEN, FN1, ATM, TP53BP1, POLR2A, SMAD4, and BRCA1 are most frequently mutated in both CESC and HNSCC. However, HRAS, TLR4, EGF, HGF, STAT1 and PTK2 are most affected in CESC and KRAS, AR, MAPK1, CUL1 and ESR1 are specific in HNSCC. Furthermore, we also demarcated essential targets based on the copy number gain and loss. PIK3CA, CCND1, RFC4, KAT5, MYC, PTK2, EGFR, and ERBB2 shows significant copy number gain proportion in CESC and HNSCC cases. Likewise, FN1, H2AFX, CHEK1, ATM, SUMO1 were marked for the substantial copy number loss in both the carcinoma and CUL1, EZH2, NRAS are unique to the HNSCC. Fibronectin 1 (FN1) could be a potential target to fight against cancer progression. FN1 is known to be involved in cell adhesion, host defense and metastasis. FN1 binds to the anastellin to form a super-fibronectin, which inhibits tumor growth, angiogenesis and metastasis. Additionally, we also proposed potential drug repurposing candidates like Dactolisib, Pilaralisib, Defactinib, Dacomitinib, Panitumumab, etc. We foresee that, this work would aid in the process of understanding HPV oncogenesis, cellular mechanisms and provide assistance towards biomarker and drug discovery to combat against HPVs.

HR- and LR-HPV types are involved in the HPV-associated diseases. Therefore, priority is given to them in terms of vaccine and therapeutic development against them. Several efforts were made to prevent HPV infection mediated diseases by employing prophylactic and immunotherapeutic vaccines. Currently, there are 3 approved virus like particle (VLP) based vaccines namely Merck's Gardasil®, a quadrivalent vaccine (HPV-6, 11, 16 and 18), GlaxoSmithKline's Cervarix®, a bivalent vaccine (HPV-16 and 18) and Gardasil®9, a nonavalent vaccine (HPV-6, 11, 16, 18, 31, 33, 45, 52 and 58) are available to protect against HPVs. However, it is reported that these are ineffective to eliminate established infections. Therefore, there is need to explore

alternative therapeutic candidates. We have developed an integrative platform; HPVomics dedicated to the HPVs potential therapeutic regimens and epitopes targeting all HPV proteins including oncoproteins E6, E7 and E5. For this, we have focused on eighteen HR-HPVs (16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 73, 82) and eleven LR-HPVs (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81).It provides therapeutically imperative elements, i.e., siRNAs, sgRNAs, anti-viral peptides, vaccine epitopes (such as IEDB epitopes, MHC-I binders, MHC-II binders, B cell and CTL epitopes) etc. Simultaneously, it also comprises whole genome sequences and annotation of all HPVs (~180) in tabular manner with searching and filtering capabilities. Additionally, it also offers interactive genome browser to visualize genomic and regulatory components powered by JBrowse. Moreover, we have also developed an integrated support vector machine (SVM) based computational algorithm "HPVepi" for the prediction of HPV epitome. We hope that HPVomics (http://bioinfo.imtech.res.in/manojk/hpvomics/) will assist the scientific community engaged in HPV research and help in subsequent crafting of therapeutic and vaccine strategies.

Next generation sequencing (NGS) provides great opportunity to study and explore viruses. Genome assembly is one of the crucial steps in the NGS data analyses. Series of distinct assemblers have been developed with the advancement in sequencing technologies. Various studies have reported the evaluation of these assembly tools on different datasets; however, these lack data from viral origin. Thus, we have evaluated and compared the performance of 8 *de novo* assemblers, i.e., SOAPdenovo, Velvet, ABySS, IDBA, SPAdes, Edena, IVA and VICUNA on the different real viral NGS datasets distinct Illumina (GAII, GAIIx, Hiseq, Miseq, Nextseq) platforms. The data belongs to the different viruses, i.e., HIV, HBV, Rhinovirus, DENV, WNV, Influenza virus, HHV, HPV and SARS-CoV-2. Performance matrices such as assembly lengths, N50, NG50, NA50, NGA50, largest contig length, contig numbers, genome fraction percentage, mis-assemblies etc. were analyzed. Two assemblers, i.e., SPAdes and IDBA performed best among all, followed by ABySS and VICUNA. Our study recommends these assemblers for the viral genome assembly. Additionally, we have also observed that the existing assemblers are inconsistent and primarily perform poor on the assembly of retro-viral data.

Viruses are most abundant and widely distributed biological bodies on earth. However, the large portion is still unknown and undiscovered. Viral metagenomics offer promising tool to explore an unprecedented diversity of viruses from distinct ecosystems. For this, different resources and computational methods have been developed to explore virome. However, distinct challenges and gaps still exist that needs to be addressed. For example, some algorithms are specific for the environmental or clinical data, lack quality control step, limited to single sequencing platform data, require computational expertise, etc. Further, existing pipelines mainly perform host subtraction step, which may end-up with loss of valuable signals like endogenous viral elements (EVEs). We have developed, VIRpipe; an integrated ready-to-use pipeline for rapid virus identification from the clinical and environmental metagenomic samples using raw data from both Illumina as well as Nanopore platform. We have utilized the sequential space reduction approach for comprehensive, sensitive yet quick viral identification and discovery. Additionally, VIRpipe can also identify Human endogenous retroviruses in clinical samples. The complete code and online version with reference data and documents is freely available at https://bioinfo.imtech.res.in/manojk/virpipe/.

Future Implications

In the current research work, we have developed distinct resources and provide analysis relevant to the HPVs, oncogenicity and viral NGS. HPVbase resource provides a knowledgebase for the potential biomarkers associated with HPV carcinomas. This includes viral integration and breakpoint events, HPVs methylation patterns and HPV mediated aberrant expression of distinct host microRNAs (miRNAs). These events would be useful to distinguish disease progression. Likewise, we also provide potential target genes and drug repurposing candidates (molecules) that could be explored to combat HPV infection and halt cancer progression. Moreover, there is still need for the effective HPV therapeutic regimen. Our resource HPVomics provides alternative therapeutics, which will be useful in the development of anti-virals against HPVs. Further, NGS provide great prospect to explore viral community and in diagnostics. We have proposed assemblers useful for the genome assembly of viral NGS data. Moreover, we have also developed ready-to-use computational pipeline for rapid identification of viruses from both environmental as well as clinical samples. This can be used to analyze metagenomic data from both the platforms, i.e., Illumina and

Nanopore. Simultaneously, this can also be utilized for the detection of endogenous viral elements (EVEs) from the clinical data.

Future Directions

With the time and advancement in technologies, new biological data used to be generated in every field including in HPV mediated diseases. Therefore, we would like to update HPVbase resource with new integration events, methylation patterns, abrupt miRNA expression and find new insights. Further, clinically important data regarding HPV/Host mutations/variations in cancers could also be incorporated in the future version. Likewise, from the analysis of HPV-associated genes, we have identified some of the potential targets and drug repurposing candidates for the HPV-infections and cancer. These potential drug molecules could be tested in experimental settings for further developments. Similarly, as we have developed a comprehensive resource HPVomics, harbouring potential therapeutic candidates like siRNAs, sgRNAs, etc. and epitopes for high-risk and low-risk HPVs. In future, some of the efficient candidates could be explored in wet-lab to inhibit HPV infection. We would also like to update this resource to add new HPV types or therapeutic category. Furthermore, we would also like to enhance the computational viral metagenomic analysis pipeline (VIRpipe). Some modules could be incorporated to enrich the pipeline like for reconstruction of complete or near complete viral genomes, automatic updating of sequence references, enrichment in characterization of sequences with low-abundance, variant detection, etc. Moreover, implementation of pipeline on the cloud clusters could be done for high throughput.

References

References

(2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378-384.

Abeles, S.R., and Pride, D.T. (2014). Molecular bases and role of viruses in the human microbiome. J Mol Biol *426*, 3892-3906.

Afiahayati, Sato, K., and Sakakibara, Y. (2013). An extended genovo metagenomic assembler by incorporating paired-end information. PeerJ *1*, e196.

Afiahayati, Sato, K., and Sakakibara, Y. (2015). MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res *22*, 69-77.

Ajami, N.J., Wong, M.C., Ross, M.C., Lloyd, R.E., and Petrosino, J.F. (2018). Maximal viral information recovery from sequence data using VirMAP. Nat Commun *9*, 3205.

Akagi, K., Li, J., Broutian, T.R., Padilla-Nash, H., Xiao, W., Jiang, B., Rocco, J.W., Teknos, T.N., Kumar, B., Wangsa, D., *et al.* (2014). Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res *24*, 185-199.

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res *40*, e126.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Amgarten, D., Braga, L.P.P., da Silva, A.M., and Setubal, J.C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. Front Genet *9*, 304.

Amoreira, C., Hindermann, W., and Grunau, C. (2003). An improved version of the DNA Methylation database (MethDB). Nucleic Acids Res *31*, 75-77.

Amoroso, R., Fitzsimmons, L., Thomas, W.A., Kelly, G.L., Rowe, M., and Bell, A.I. (2011). Quantitative studies of Epstein-Barr virus-encoded microRNAs provide novel insights into their regulation. J Virol *85*, 996-1010.

Anderson, N.G., Gerin, J.L., and Anderson, N.L. (2003). Global screening for human viral pathogens. Emerg Infect Dis *9*, 768-774.

Andersson, S., Alemi, M., Rylander, E., Strand, A., Larsson, B., Sallstrom, J., and Wilander, E. (2000). Uneven distribution of HPV 16 E6 prototype and variant (L83V) oncoprotein in cervical neoplastic lesions. Br J Cancer *83*, 307-310.

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., and Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health *8*, e191-e203.

Arlotta, P., and Macklis, J.D. (2005). Archeo-cell biology: carbon dating is not just for pots and dinosaurs. Cell *122*, 4-6.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res *44*, W16-21.

Asadurian, Y., Kurilin, H., Lichtig, H., Jackman, A., Gonen, P., Tommasino, M., Zehbe, I., and Sherman, L. (2007). Activities of human papillomavirus 16 E6 natural variants in human keratinocytes. J Med Virol *79*, 1751-1760.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Ayling, M., Clark, M.D., and Leggett, R.M. (2020). New approaches for metagenome assembly with short reads. Brief Bioinform *21*, 584-594.

Baaijens, J.A., Aabidine, A.Z.E., Rivals, E., and Schönhuth, A. (2017). De novo assembly of viral quasispecies using overlap graphs. Genome Res *27*, 835-848.

Badal, S., Badal, V., Calleja-Macias, I.E., Kalantari, M., Chuang, L.S., Li, B.F., and Bernard, H.U. (2004). The human papillomavirus-18 genome is efficiently targeted by cellular DNA methylation. Virology *324*, 483-492.

Badal, V., Chuang, L.S., Tan, E.H., Badal, S., Villa, L.L., Wheeler, C.M., Li, B.F., and Bernard, H.U. (2003). CpG methylation of human papillomavirus type 16 DNA in cervical cancer cell lines and in clinical specimens: genomic hypomethylation correlates with carcinogenic progression. J Virol *77*, 6227-6234.

Baheti, S., Tang, X., O'Brien, D.R., Chia, N., Roberts, L.R., Nelson, H., Boughey, J.C., Wang, L., Goetz, M.P., Kocher, J.A., *et al.* (2018). HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data. BMC Bioinformatics *19*, 271.

Balcazar, J.L. (2014). Bacteriophages as vehicles for antibiotic resistance genes in the environment. PLoS Pathog *10*, e1004219.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455-477.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y.Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet *56*, 406-414.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691-1692.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Barthelson, R., McFarlin, A.J., Rounsley, S.D., and Young, S. (2011). Plantagora: modeling whole genome sequencing and assembly of plant genomes. PLoS One *6*, e28436.

Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S., and Palù, G. (2013). Next-generation sequencing technologies in diagnostic virology. J Clin Virol *58*, 346-350.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., and Palu, G. (2011a). Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci *12*, 7861-7884.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., and Palù, G. (2011b). Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci *12*, 7861-7884.

Bauman, Y., Nachmani, D., Vitenshtein, A., Tsukerman, P., Drayman, N., Stern-Ginossar, N., Lankry, D., Gruda, R., and Mandelboim, O. (2011). An identical miRNA of the human JC and BK polyoma viruses targets the stress-induced ligand ULBP3 to escape immune elimination. Cell Host Microbe *9*, 93-102.

Beerenwinkel, N., Günthard, H.F., Roth, V., and Metzner, K.J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol *3*, 329.

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol *33*, 623-630.

Bernard, H.U., Calleja-Macias, I.E., and Dunn, S.T. (2006). Genome variation of human papillomavirus types: phylogenetic and medical implications. Int J Cancer *118*, 1071-1076.

Berumen, J., Ordonez, R.M., Lazcano, E., Salmeron, J., Galvan, S.C., Estrada, R.A., Yunes, E., Garcia-Carranca, A., Gonzalez-Lira, G., and Madrigal-de la Campa, A. (2001). Asian-American variants of human papillomavirus 16 and risk for cervical cancer: a case-control study. J Natl Cancer Inst *93*, 1325-1330.

Berzofsky, J.A., Ahlers, J.D., Janik, J., Morris, J., Oh, S., Terabe, M., and Belyakov, I.M. (2004). Progress on new vaccine strategies against chronic viral infections. J Clin Invest *114*, 450-462.

Besecker, M.I., Harden, M.E., Li, G., Wang, X.J., and Griffiths, A. (2009). Discovery of herpes B virus-encoded microRNAs. J Virol *83*, 3413-3416.

Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. Nucleic Acids Res *36*, D149-153.

Bhasin, M., and Raghava, G.P. (2004). Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine *22*, 3195-3204.

Bhattacharjee, B., and Sengupta, S. (2006a). CpG methylation of HPV 16 LCR at E2 binding site proximal to P97 is associated with cervical cancer in presence of intact E2. Virology *354*, 280-285.

Bhattacharjee, B., and Sengupta, S. (2006b). HPV16 E2 gene disruption and polymorphisms of E2 and LCR: some significant associations with cervical cancer in Indian women. Gynecol Oncol *100*, 372-378.

Bhuvaneshwar, K., Song, L., Madhavan, S., and Gusev, Y. (2018). viGEN: An Open Source Pipeline for the Detection and Quantification of Viral RNA in Human Tumors. Front Microbiol *9*, 1172.

Bielewicz, D., Dolata, J., Zielezinski, A., Alaba, S., Szarzynska, B., Szczesniak, M.W., Jarmolowski, A., Szweykowska-Kulinska, Z., and Karlowski, W.M. (2012). mirEX: a platform for comparative exploration of plant pri-miRNA expression data. Nucleic Acids Res *40*, D191-197.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol *13*, R122.

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B.L., and Sullivan, M.B. (2017). iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J *11*, 7-14.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Bolognini, D., Bartalucci, N., Mingrino, A., Vannucchi, A.M., and Magi, A. (2019). NanoR: A user-friendly R package to analyze and compare nanopore sequencing data. PLoS One *14*, e0216471.

Borozan, I., Wilson, S., Blanchette, P., Laflamme, P., Watt, S.N., Krzyzanowski, P.M., Sircoulomb, F., Rottapel, R., Branton, P.E., and Ferretti, V. (2012). CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics *13*, 206.

Bosch, F.X., Lorincz, A., Munoz, N., Meijer, C.J., and Shah, K.V. (2002). The causal relation between human papillomavirus and cervical cancer. J Clin Pathol *55*, 244-265.

Bose, M., and Barber, R.D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biol *6*, 223-227.

Boulet, G.A., Horvath, C.A., Berghmans, S., and Bogers, J. (2008). Human papillomavirus in cervical cancer screening: important role as biomarker. Cancer Epidemiol Biomarkers Prev *17*, 810-817.

Bourgault Villada, I., Beneton, N., Bony, C., Connan, F., Monsonego, J., Bianchi, A., Saiag, P., Levy, J.P., Guillet, J.G., and Choppin, J. (2000). Identification in humans of HPV-16 E6 and E7 protein epitopes recognized by cytolytic T lymphocytes in association with HLA-B18 and determination of the HLA-B18-specific binding motif. Eur J Immunol *30*, 2281-2289.

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., and Tyson, G.W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol *9*, e1003031.

Brand, T.M., Hartmann, S., Bhola, N.E., Peyser, N.D., Li, H., Zeng, Y., Isaacson Wechsler, E., Ranall, M.V., Bandyopadhyay, S., Duvvuri, U., *et al.* (2017). Human Papillomavirus Regulates HER3 Expression in Head and Neck Cancer: Implications for Targeted HER3 Therapy in HPV(+) Patients. Clin Cancer Res *23*, 3072-3083.

Brandsma, J.L., Harigopal, M., Kiviat, N.B., Sun, Y., Deng, Y., Zelterman, D., Lizardi, P.M., Shabanova, V.S., Levi, A., Yaping, T., *et al.* (2014). Methylation of Twelve CpGs in Human Papillomavirus Type 16 (HPV16) as an Informative Biomarker for the Triage of Women Positive for HPV16 Infection. Cancer Prev Res (Phila).

Brandsma, J.L., Sun, Y., Lizardi, P.M., Tuck, D.P., Zelterman, D., Haines, G.K., 3rd, Martel, M., Harigopal, M., Schofield, K., and Neapolitano, M. (2009). Distinct human papillomavirus type 16 methylomes in cervical cells at different stages of premalignancy. Virology *389*, 100-107.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin *68*, 394-424.

Brianti, P., De Flammineis, E., and Mercuri, S.R. (2017). Review of HPV-related diseases and cancers. New Microbiol *40*, 80-85.

Briese, T., Kapoor, A., Mishra, N., Jain, K., Kumar, A., Jabado, O.J., and Lipkin, W.I. (2015). Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. MBio *6*, e01491-01415.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res *43*, D571-577.

Brister, J.R., Bao, Y., Zhdanov, S.A., Ostapchuck, Y., Chetvernin, V., Kiryutin, B., Zaslavsky, L., Kimelman, M., and Tatusova, T.A. (2014). Virus Variation Resource--recent updates and future directions. Nucleic Acids Res *42*, D660-665.

Bryant, D., Tristram, A., Liloglou, T., Hibbitts, S., Fiander, A., and Powell, N. (2014). Quantitative measurement of Human Papillomavirus type 16 L1/L2 DNA methylation correlates with cervical disease grade. J Clin Virol *59*, 24-29.

Burk, R.D., Harari, A., and Chen, Z. (2013). Human papillomavirus genome variants. Virology *445*, 232-243.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res *18*, 810-820.

Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., Siegel, M., Benson, G., Crandall, K.A., and Johnson, W.E. (2014). Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinformatics *15*, 262.

Cadwell, K. (2015). The virome in host health and disease. Immunity *42*, 805-813.

Cai, H.B., Chen, C.C., and Ding, X.H. (2010). Human papillomavirus type 16 E6 gene variations in Chinese population. Eur J Surg Oncol *36*, 160-163.

Calin, G.A., and Croce, C.M. (2006). MicroRNA signatures in human cancers. Nat Rev Cancer *6*, 857-866.

Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., *et al.* (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A *101*, 2999-3004.

Calis, J.J., Maybeno, M., Greenbaum, J.A., Weiskopf, D., De Silva, A.D., Sette, A., Kesmir, C., and Peters, B. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. PLoS Comput Biol *9*, e1003266.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Campbell, I.M., Gambin, T., Dittwald, P., Beck, C.R., Shuvarikov, A., Hixson, P., Patel, A., Gambin, A., Shaw, C.A., Rosenfeld, J.A., *et al.* (2014). Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. BMC Biol *12*, 74.

Cannizzaro, L.A., Durst, M., Mendez, M.J., Hecht, B.K., and Hecht, F. (1988). Regional chromosome localization of human papillomavirus integration sites near fragile sites, oncogenes, and cancer chromosome breakpoints. Cancer Genet Cytogenet *33*, 93-98.

Capobianchi, M.R., Giombini, E., and Rozera, G. (2013). Next-generation sequencing technology in clinical virology. Clin Microbiol Infect *19*, 15-22.

Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O., and Mazet, J.A.K. (2018). The Global Virome Project. Science *359*, 872-874.

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., *et al.* (2019). Open Targets Platform: new developments and updates two years on. Nucleic Acids Res *47*, D1056-d1065.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., *et al.* (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov *2*, 401-404.

Chabeda, A., Yanez, R.J.R., Lamprecht, R., Meyers, A.E., Rybicki, E.P., and Hitzeroth, II (2018). Therapeutic vaccines for high-risk HPV-associated diseases. Papillomavirus Res *5*, 46-58.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics *13*, 238.

Chakrabarti, O., Veeraraghavalu, K., Tergaonkar, V., Liu, Y., Androphy, E.J., Stanley, M.A., and Krishna, S. (2004). Human papillomavirus type 16 E6 amino acid 83 variants enhance E6-mediated MAPK signaling and differentially regulate tumorigenesis by notch signaling and oncogenic Ras. J Virol *78*, 5934-5945.

Chan, P.K., Lam, C.W., Cheung, T.H., Li, W.W., Lo, K.W., Chan, M.Y., Cheung, J.L., Xu, L.Y., and Cheng, A.F. (2002). Human papillomavirus type 16 intratypic variant infection and risk for cervical neoplasia in southern China. J Infect Dis *186*, 696-700.

Chandrani, P., Kulkarni, V., Iyer, P., Upadhyay, P., Chaubal, R., Das, P., Mulherkar, R., Singh, R., and Dutt, A. (2015). NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. Br J Cancer *112*, 1958-1965.

Chang, J.L., Tsao, Y.P., Liu, D.W., Huang, S.J., Lee, W.H., and Chen, S.L. (2001). The expression of HPV-16 E5 protein in squamous neoplastic changes in the uterine cervix. J Biomed Sci *8*, 206-213.

Chang, J.T., Kuo, T.F., Chen, Y.J., Chiu, C.C., Lu, Y.C., Li, H.F., Shen, C.R., and Cheng, A.J. (2010). Highly potent and specific siRNAs against E6 or E7 genes of HPV16- or HPV18-infected cervical cancers. Cancer Gene Ther *17*, 827-836.

Chang, S., Zhang, J., Liao, X., Zhu, X., Wang, D., Zhu, J., Feng, T., Zhu, B., Gao, G.F., Wang, J., *et al.* (2007). Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. Nucleic Acids Res *35*, D376-380.

Chansaenroj, J., Theamboonlers, A., Junyangdikul, P., Swangvaree, S., Karalak, A., and Poovorawan, Y. (2012). Whole genome analysis of human papillomavirus type 16 multiple infection in cervical cancer patients. Asian Pac J Cancer Prev *13*, 599-606.

Cheloufi, S., Dos Santos, C.O., Chong, M.M., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584-589.

Chen, C.J., Kincaid, R.P., Seo, G.J., Bennett, M.D., and Sullivan, C.S. (2011). Insights into Polyomaviridae microRNA function derived from study of the bandicoot papillomatosis carcinomatosis viruses. J Virol *85*, 4487-4500.

Chen, J., Huang, J., and Sun, Y. (2019). TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. BMC Bioinformatics *20*, 305.

Chen, T.W., Gan, R.R., Wu, T.H., Lin, W.C., and Tang, P. (2012). VIP DB--a viral protein domain usage and distribution database. Genomics *100*, 149-156.

Chen, Y., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N., and Su, X. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics *29*, 266-267.

Chen, Z., de Freitas, L.B., and Burk, R.D. (2015). Evolution and classification of oncogenic human papillomavirus types and variants associated with cervical cancer. Methods Mol Biol *1249*, 3-26.

Chen, Z., Storthz, K.A., and Shillitoe, E.J. (1997). Mutations in the long control region of human papillomavirus DNA in oral cancer cells, and their functional consequences. Cancer Res *57*, 1614-1619.

Chen, Z., Terai, M., Fu, L., Herrero, R., DeSalle, R., and Burk, R.D. (2005). Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. J Virol *79*, 7014-7023.

Cheng, M.A., Farmer, E., Huang, C., Lin, J., Hung, C.F., and Wu, T.C. (2018). Therapeutic DNA Vaccines for Human Papillomavirus and Associated Diseases. Hum Gene Ther *29*, 971-996.

Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T., and Lin, C.Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol *8 Suppl 4*, S11.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., *et al.* (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods *10*, 563-569.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., *et al.* (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods *13*, 1050-1054.

Cho, S., Jang, I., Jun, Y., Yoon, S., Ko, M., Kwon, Y., Choi, I., Chang, H., Ryu, D., Lee, B., *et al.* (2013). MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic Acids Res *41*, D252-257.

Choo, K.B., Lee, H.H., Pan, C.C., Wu, S.M., Liew, L.N., Cheung, W.F., and Han, S.H. (1988). Sequence duplication and internal deletion in the integrated human papillomavirus type 16 genome cloned from a cervical carcinoma. J Virol *62*, 1659-1666.

Choo, K.B., Wang, T.S., and Huang, C.J. (2000). Analysis of relative binding affinity of E7-pRB of human papillomavirus 16 clinical variants using the yeast two-hybrid system. J Med Virol *61*, 298-302.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., *et al.* (2010). A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science *328*, 1694-1698.

Clarke, M.A., Wentzensen, N., Mirabello, L., Ghosh, A., Wacholder, S., Harari, A., Lorincz, A., Schiffman, M., and Burk, R.D. (2012). Human papillomavirus DNA methylation as a potential biomarker for cervical cancer. Cancer Epidemiol Biomarkers Prev *21*, 2125-2137.

Clifford, G., Franceschi, S., Diaz, M., Munoz, N., and Villa, L.L. (2006). Chapter 3: HPV type-distribution in women with and without cervical neoplastic diseases. Vaccine *24 Suppl 3*, S3/26-34.

Clokie, M.R., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011). Phages in nature. Bacteriophage *1*, 31-45.

Cogliano, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., and El Ghissassi, F. (2005). Carcinogenicity of human papillomaviruses. Lancet Oncol *6*, 204.

Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., Dehne-Garcia, A., Geourjon, C., Bettler, E., Hulo, C., *et al.* (2007). euHCVdb: the European hepatitis C virus database. Nucleic Acids Res *35*, D363-366.

Cornet, I., Gheit, T., Franceschi, S., Vignat, J., Burk, R.D., Sylla, B.S., Tommasino, M., and Clifford, G.M. (2012). Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR. J Virol *86*, 6855-6861.

Costa, R.L., Boroni, M., and Soares, M.A. (2018). Distinct co-expression networks using multi-omic data reveal novel interventional targets in HPV-positive and negative head-and-neck squamous cell cancer. Sci Rep *8*, 15254.

Crosbie, E.J., Einstein, M.H., Franceschi, S., and Kitchener, H.C. (2013). Human papillomavirus and cervical cancer. Lancet *382*, 889-899.

D'Souza, G., and Dempsey, A. (2011). The role of HPV in head and neck cancer and review of the HPV vaccine. Prev Med *53 Suppl 1*, S5-s11.

Dadar, M., Chakraborty, S., Dhama, K., Prasad, M., Khandia, R., Hassan, S., Munjal, A., Tiwari, R., Karthik, K., Kumar, D., *et al.* (2018). Advances in Designing and Developing Vaccines, Drugs and Therapeutic Approaches to Counter Human Papilloma Virus. Front Immunol *9*, 2478.

Dar, S.A., Gupta, A.K., Thakur, A., and Kumar, M. (2016). SMEpred workbench: A web server for predicting efficacy of chemicallymodified siRNAs. RNA Biol *13*, 1144-1151.

de Araujo Souza, P.S., Maciag, P.C., Ribeiro, K.B., Petzl-Erler, M.L., Franco, E.L., and Villa, L.L. (2008). Interaction between polymorphisms of the human leukocyte antigen and HPV-16 variants on the risk of invasive cervical cancer. BMC Cancer *8*, 246.

de Boer, M.A., Peters, L.A., Aziz, M.F., Siregar, B., Cornain, S., Vrede, M.A., Jordanova, E.S., Kolkman-Uljee, S., and Fleuren, G.J. (2004). Human papillomavirus type 16 E6, E7, and L1 variants in cervical cancer in Indonesia, Suriname, and The Netherlands. Gynecol Oncol *94*, 488-494.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666-2669.

de Freitas, A.C., Coimbra, E.C., and Leitão Mda, C. (2014). Molecular targets of HPV oncoproteins: potential biomarkers for cervical carcinogenesis. Biochim Biophys Acta *1845*, 91-103.

de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G.M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health *8*, e180-e190.

de Martel, C., Plummer, M., Vignat, J., and Franceschi, S. (2017). Worldwide burden of cancer attributable to HPV by site, country and HPV type. Int J Cancer *141*, 664-670.

de Sanjose, S., Quint, W.G., Alemany, L., Geraets, D.T., Klaustermeier, J.E., Lloveras, B., Tous, S., Felix, A., Bravo, L.E., Shin, H.R., *et al.* (2010). Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. Lancet Oncol *11*, 1048-1056.

de Sousa, A.L., Maués, D., Lobato, A., Franco, E.F., Pinheiro, K., Araújo, F., Pantoja, Y., da Costa da Silva, A.L., Morais, J., and Ramos, R.T.J. (2018). PhageWeb -

Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes. Front Genet *9*, 644.

de Villiers, E.M. (2013). Cross-roads in the classification of papillomaviruses. Virology *445*, 2-10.

de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U., and zur Hausen, H. (2004). Classification of papillomaviruses. Virology *324*, 17-27.

Delwart, E.L. (2007). Viral metagenomics. Rev Med Virol *17*, 115-131.

Descamps, D., Hardt, K., Spiessens, B., Izurieta, P., Verstraeten, T., Breuer, T., and Dubin, G. (2009). Safety of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine for cervical cancer prevention: a pooled analysis of 11 clinical trials. Hum Vaccin *5*, 332-340.

Devaraj, K., Gillison, M.L., and Wu, T.C. (2003). Development of HPV vaccines for HPV-associated head and neck squamous cell carcinoma. Crit Rev Oral Biol Med *14*, 345-362.

Dey, L., and Mukhopadhyay, A. (2017). DenvInt: A database of protein-protein interactions between dengue virus and its hosts. PLoS Negl Trop Dis *11*, e0005879.

Dhanda, S.K., Chaudhary, K., Gupta, S., Brahmachari, S.K., and Raghava, G.P. (2016). A web-based resource for designing therapeutics against Ebola Virus. Sci Rep *6*, 24782.

Dhanda, S.K., Karosiene, E., Edwards, L., Grifoni, A., Paul, S., Andreatta, M., Weiskopf, D., Sidney, J., Nielsen, M., Peters, B., *et al.* (2018). Predicting HLA CD4 Immunogenicity in Human Populations. Front Immunol *9*, 1369.

Diederichs, S., and Haber, D.A. (2007). Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. Cell *131*, 1097-1108.

DiMaio, D., and Petti, L.M. (2013). The E5 proteins. Virology *445*, 99-114.

Ding, D.C., Chiang, M.H., Lai, H.C., Hsiung, C.A., Hsieh, C.Y., and Chu, T.Y. (2009). Methylation of the long control region of HPV16 is related to the severity of cervical neoplasia. Eur J Obstet Gynecol Reprod Biol *147*, 215-220.

Ding, T., Wang, X., Ye, F., Cheng, X., Lu, W., and Xie, X. (2010). Distribution of human papillomavirus 16 E6/E7 variants in cervical cancer and intraepithelial neoplasia in Chinese women. Int J Gynecol Cancer *20*, 1391-1398.

Doerfler, W. (2005). On the biological significance of DNA methylation. Biochemistry (Mosc) *70*, 505-524.

Dong, X.P., Stubenrauch, F., Beyer-Finkler, E., and Pfister, H. (1994). Prevalence of deletions of YY1-binding sites in episomal HPV 16 DNA from cervical cancers. Int J Cancer *58*, 803-808.

Doorbar, J. (2006). Molecular biology of human papillomavirus infection and cervical cancer. Clin Sci (Lond) *110*, 525-541.

Doorbar, J., Egawa, N., Griffin, H., Kranjec, C., and Murakami, I. (2015). Human papillomavirus molecular biology and disease association. Rev Med Virol *25 Suppl 1*, 2-23.

Doorbar, J., Quint, W., Banks, L., Bravo, I.G., Stoler, M., Broker, T.R., and Stanley, M.A. (2012). The biology and life-cycle of human papillomaviruses. Vaccine *30 Suppl 5*, F55-70.

Duensing, S., and Munger, K. (2002). The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability. Cancer Res *62*, 7075-7082.

Durkin, S.G., and Glover, T.W. (2007). Chromosome fragile sites. Annu Rev Genet *41*, 169-192.

Dyson, N., Howley, P.M., Munger, K., and Harlow, E. (1989). The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. Science *243*, 934-937.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M., *et al.* (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res *21*, 2224-2241.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. Nat Rev Microbiol *3*, 504-510.

Einstein, M.H., Kadish, A.S., Burk, R.D., Kim, M.Y., Wadler, S., Streicher, H., Goldberg, G.L., and Runowicz, C.D. (2007). Heat shock fusion protein-based immunotherapy for treatment of cervical intraepithelial neoplasia III. Gynecol Oncol *106*, 453-460.

Einstein, M.H., Leanza, S., Chiu, L.G., Schlecht, N.F., Goldberg, G.L., Steinberg, B.M., and Burk, R.D. (2009). Genetic variants in TAP are associated with high-grade cervical neoplasia. Clin Cancer Res *15*, 1019-1023.

Elbehery, A.H.A., Feichtmayer, J., Singh, D., Griebler, C., and Deng, L. (2018). The Human Virome Protein Cluster Database (HVPC): A Human Viral Metagenomic Database for Diversity and Function Annotation. Front Microbiol *9*, 1110.

Engels, B.M., and Hutvagner, G. (2006). Principles and effects of microRNA-mediated post-transcriptional gene regulation. Oncogene *25*, 6163-6169.

Eschle, D., Durst, M., ter Meulen, J., Luande, J., Eberhardt, H.C., Pawlita, M., and Gissmann, L. (1992). Geographical dependence of sequence variation in the E7 gene of human papillomavirus type 16. J Gen Virol *73 (Pt 7)*, 1829-1832.

Esquela-Kerscher, A., and Slack, F.J. (2006). Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer *6*, 259-269.

Facciuto, F., Bugnon Valdano, M., Marziali, F., Massimi, P., Banks, L., Cavatorta, A.L., and Gardiol, D. (2014). Human papillomavirus (HPV)-18 E6 oncoprotein interferes with the epithelial cell polarity Par3 protein. Mol Oncol *8*, 533-543.

Fancello, L., Raoult, D., and Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. Virology *434*, 162-174.

Fang, E., and Zhang, X. (2017). Identification of breast cancer hub genes and analysis of prognostic values using integrated bioinformatics analysis. Cancer Biomark *21*, 373-381.

Fang, E., Zhang, X., Wang, Q., and Wang, D. (2017). Identification of prostate cancer hub genes and therapeutic agents using bioinformatics approach. Cancer Biomark *20*, 553-561.

Farazi, T.A., Spitzer, J.I., Morozov, P., and Tuschl, T. (2011). miRNAs in human cancer. J Pathol *223*, 102-115.

Fedonin, G.G., Fantin, Y.S., Favorov, A.V., Shipulin, G.A., and Neverov, A.D. (2019). VirGenA: a reference-based assembler for variable viral genomes. Brief Bioinform *20*, 15-25.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality

worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer *136*, E359-386.

Fernandez, A.F., Rosales, C., Lopez-Nieva, P., Grana, O., Ballestar, E., Ropero, S., Espada, J., Melo, S.A., Lujambio, A., Fraga, M.F., *et al.* (2009). The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer. Genome Res *19*, 438-451.

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet *13*, 283-296.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M., Di Camillo, B., and Toppo, S. (2012). Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. Brief Bioinform *13*, 269-280.

Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E.H., Tardif, K.D., Kapusta, A., Rynearson, S., *et al.* (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol *17*, 111.

Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics *28*, 3169-3177.

Forman, D., de Martel, C., Lacey, C.J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., Vignat, J., Ferlay, J., Bray, F., Plummer, M., *et al.* (2012). Global burden of human papillomavirus and related diseases. Vaccine *30 Suppl 5*, F12-23.

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Rühlemann, M., Kraemer, L., Mucha, S., Wienbrandt, L., Stanulla, M., and Franke, A. (2015). Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. Sci Rep *5*, 11534.

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A., Pariente, K., Segondy, M., Burguiere, A., Manuguerra, J.C., *et al.* (2012a). Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One *7*, e38499.

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A., Pariente, K., Segondy, M., Burguière, A., Manuguerra, J.C., *et al.* (2012b). Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One *7*, e38499.

Fouts, D.E. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res *34*, 5839-5851.

Frazer, I.H. (2004). Prevention of cervical cancer through papillomavirus vaccination. Nat Rev Immunol *4*, 46-54.

Fujinaga, Y., Okazawa, K., Nishikawa, A., Yamakawa, Y., Fukushima, M., Kato, I., and Fujinaga, K. (1994). Sequence variation of human papillomavirus type 16 E7 in preinvasive and invasive cervical neoplasias. Virus Genes *9*, 85-92.

Fuks, F. (2005). DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev *15*, 490-495.

Galloway, D.A. (2003). Papillomavirus vaccines in clinical trials. Lancet Infect Dis *3*, 469-475.

Garcia-Hernandez, E., Gonzalez-Sanchez, J.L., Andrade-Manzano, A., Contreras, M.L., Padilla, S., Guzman, C.C., Jimenez, R., Reyes, L., Morosoli, G., Verde, M.L., *et al.* (2006). Regression of papilloma high-grade lesions (CIN 2 and CIN 3) is stimulated by therapeutic vaccination with MVA E2 recombinant vaccine. Cancer Gene Ther *13*, 592-597.

Garcia-Vallve, S., Alonso, A., and Bravo, I.G. (2005). Papillomaviruses: different genes have different histories. Trends Microbiol *13*, 514-521.

Garretto, A., Hatzopoulos, T., and Putonti, C. (2019). virMine: automated detection of viral sequences from complex metagenomic samples. PeerJ *7*, e6695.

Gaykalova, D.A., Mambo, E., Choudhary, A., Houghton, J., Buddavarapu, K., Sanford, T., Darden, W., Adai, A., Hadd, A., Latham, G., *et al.* (2014). Novel insight into mutational landscape of head and neck squamous cell carcinoma. PLoS One *9*, e93102.

Geisen, S., Barturen, G., Alganza, A.M., Hackenberg, M., and Oliver, J.L. (2014). NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. Nucleic Acids Res *42*, D53-59.

Gerlach, D., Kriventseva, E.V., Rahman, N., Vejnar, C.E., and Zdobnov, E.M. (2009). miROrtho: computational survey of microRNA genes. Nucleic Acids Res *37*, D111-117.

Ghosh, T.S., Mohammed, M.H., Komanduri, D., and Mande, S.S. (2011). ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. Bioinformation *6*, 91-94.

Giannoudis, A., Duin, M., Snijders, P.J., and Herrington, C.S. (2001). Variation in the E2-binding domain of HPV 16 is associated with high-grade squamous intraepithelial lesions of the cervix. Br J Cancer *84*, 1058-1063.

Giannoudis, A., and Herrington, C.S. (2001). Human papillomavirus variants and squamous neoplasia of the cervix. J Pathol *193*, 295-302.

Gillison, M.L., Koch, W.M., Capone, R.B., Spafford, M., Westra, W.H., Wu, L., Zahurak, M.L., Daniel, R.W., Viglione, M., Symer, D.E., *et al.* (2000). Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. J Natl Cancer Inst *92*, 709-720.

Gillison, M.L., and Shah, K.V. (2003). Chapter 9: Role of mucosal human papillomavirus in nongenital cancers. J Natl Cancer Inst Monogr, 57-65.

Gocze, K., Gombos, K., Juhasz, K., Kovacs, K., Kajtar, B., Benczik, M., Gocze, P., Patczai, B., Arany, I., and Ember, I. (2013). Unique microRNA expression profiles in cervical cancer. Anticancer Res *33*, 2561-2567.

Gomez-Gomez, Y., Organista-Nava, J., and Gariglio, P. (2013). Deregulation of the miRNAs expression in cervical cancer: human papillomavirus implications. Biomed Res Int *2013*, 407052.

Govan, V.A. (2008). A novel vaccine for cervical cancer: quadrivalent human papillomavirus (types 6, 11, 16 and 18) recombinant vaccine (Gardasil). Ther Clin Risk Manag *4*, 65-70.

Grabowska, A.K., and Riemer, A.B. (2012). The invisible enemy - how human papillomaviruses avoid recognition and clearance by the host immune system. Open Virol J *6*, 249-256.

Grant, J.R., and Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res *36*, W181-184.

Gregor, I., Schönhuth, A., and McHardy, A.C. (2016). Snowball: strain aware gene assembly of metagenomes. Bioinformatics *32*, i649-i657.

Gregory, R.I., Chendrimada, T.P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell *123*, 631-640.

Greninger, A.L., Naccache, S.N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J.M., *et al.* (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med *7*, 99.

Grodzki, M., Besson, G., Clavel, C., Arslan, A., Franceschi, S., Birembaut, P., Tommasino, M., and Zehbe, I. (2006). Increased risk for cervical disease progression of French women infected with the human papillomavirus type 16 E6-350G variant. Cancer Epidemiol Biomarkers Prev *15*, 820-822.

Grundhoff, A., and Sullivan, C.S. (2011). Virus-encoded microRNAs. Virology *411*, 325-343.

Gu, W., An, J., Ye, P., Zhao, K.N., and Antonsson, A. (2011). Prediction of conserved microRNAs from skin and mucosal human papillomaviruses. Arch Virol *156*, 1161-1171.

Gupta, A.K., Kaur, K., Rajput, A., Dhanda, S.K., Sehgal, M., Khan, M.S., Monga, I., Dar, S.A., Singh, S., Nagpal, G., *et al.* (2016). ZikaVR: An Integrated Zika Virus Resource for Genomics, Proteomics, Phylogenetic and Therapeutic Analysis. Sci Rep *6*, 32713.

Gupta, A.K., Khan, M.S., Choudhury, S., Mukhopadhyay, A., Sakshi, Rastogi, A., Thakur, A., Kumari, P., Kaur, M., Shalu, *et al.* (2020a). CoronaVR: A Computational Resource and Analysis of Epitopes and Therapeutics for Severe Acute Respiratory Syndrome Coronavirus-2. Front Microbiol *11*, 1858.

Gupta, A.K., Kumar, A., Rajput, A., Kaur, K., Dar, S.A., Thakur, A., Megha, K., and Kumar, M. (2020b). NipahVR: a resource of multi-targeted putative therapeutics and epitopes for the Nipah virus. Database (Oxford) *2020*.

Gupta, A.K., and Kumar, M. (2015). Landscape of HPV mediated events as potential biomarkers in diverse carcinomas. Paper presented at: VIROCON - 2015, XXIV National Conference of Indian Virological Society (Indian Virological Society).

Gupta, A.K., and Kumar, M. (2016). HPV integration associated genome-wide disruption –A functional and network analysis. Paper presented at: Proceedings of NextGen Genomics, Biology and Bioinformatics and Technologies (NGBT) International Conference (SciGenom Research Foundation (SGRF)).

Gupta, A.K., and Kumar, M. (2020). HPVomics: An integrated resource for the human papillomavirus epitome and therapeutics. Genomics.

Gupta, S., Kumar, P., and Das, B.C. (2018). HPV: Molecular pathways and targets. Curr Probl Cancer *42*, 161-174.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072-1075.

Gurumayum, S., Brahma, R., Naorem, L.D., Muthaiyan, M., Gopal, J., and Venkatesan, A. (2018). ZikaBase: An integrated ZIKV- Human Interactome Map database. Virology *514*, 203-210.

Haider, B., Ahn, T.H., Bushnell, B., Chai, J., Copeland, A., and Pan, C. (2014). Omega: an overlap-graph de novo assembler for metagenomics. Bioinformatics *30*, 2717-2722.

Hallez, S., Simon, P., Maudoux, F., Doyen, J., Noel, J.C., Beliard, A., Capelle, X., Buxant, F., Fayt, I., Lagrost, A.C., *et al.* (2004). Phase I/II trial of immunogenicity of a human papillomavirus (HPV) type 16 E7 protein-based vaccine in women with oncogenic HPV-positive cervical intraepithelial neoplasia. Cancer Immunol Immunother *53*, 642-650.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev *18*, 3016-3027.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887-901.

Hancock, G., Hellner, K., and Dorrell, L. (2018). Therapeutic HPV vaccines. Best Pract Res Clin Obstet Gynaecol *47*, 59-72.

Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodkinson, B.P., SanMiguel, A.J., Minot, S., Bushman, F.D., and Grice, E.A. (2015). The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. MBio *6*, e01578-01515.

Hao, Y., Yang, L., Galvao Neto, A., Amin, M.R., Kelly, D., Brown, S.M., Branski, R.C., and Pei, Z. (2018). HPViewer: sensitive and specific genotyping of human papillomavirus in metagenomic DNA. Bioinformatics *34*, 1986-1995.

Harper, D.M. (2009). Current prophylactic HPV vaccines and gynecologic premalignancies. Curr Opin Obstet Gynecol *21*, 457-464.

Harper, D.M., Franco, E.L., Wheeler, C., Ferris, D.G., Jenkins, D., Schuind, A., Zahaf, T., Innis, B., Naud, P., De Carvalho, N.S., *et al.* (2004). Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomised controlled trial. Lancet *364*, 1757-1765.

Hatem, A., Bozdağ, D., Toland, A.E., and Çatalyürek Ü, V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14*, 184.

Hayes, S., Mahony, J., Nauta, A., and van Sinderen, D. (2017). Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. Viruses *9*.

He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., Yang, L., Sun, Z.S., Yang, H., and Wang, J. (2008). MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res *36*, D836-841.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res *18*, 802-809.

Hildesheim, A., Herrero, R., Wacholder, S., Rodriguez, A.C., Solomon, D., Bratti, M.C., Schiller, J.T., Gonzalez, P., Dubin, G., Porras, C., *et al.* (2007). Effect of human papillomavirus 16/18 L1 viruslike particle vaccine among young women with preexisting infection: a randomized trial. JAMA *298*, 743-753.

Ho, D.W., Sze, K.M., and Ng, I.O. (2015). Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget *6*, 20959-20963.

Ho, L., Chan, S.Y., Burk, R.D., Das, B.C., Fujinaga, K., Icenogle, J.P., Kahn, T., Kiviat, N., Lancaster, W., Mavromara-Nazos, P., *et al.* (1993). The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. J Virol *67*, 6413-6423.

Ho, T., and Tzanetakis, I.E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. Virology *471-473*, 54-60.

Holmes, E.C. (2011). The evolution of endogenous viral elements. Cell Host Microbe *10*, 368-377.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J.F., Byrd, A.L., Castro-Nallar, E., Crandall, K.A., and Johnson, W.E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome *2*, 33.

Hoppe-Seyler, K., Bossler, F., Braun, J.A., Herrmann, A.L., and Hoppe-Seyler, F. (2018). The HPV E6/E7 Oncogenes: Key Factors for Viral Carcinogenesis and Therapeutic Targets. Trends Microbiol *26*, 158-168.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M., *et al.* (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature *463*, 84-87.

Hsu, P.W., Lin, L.Z., Hsu, S.D., Hsu, J.B., and Huang, H.D. (2007). ViTa: prediction of host microRNAs targets on viruses. Nucleic Acids Res *35*, D381-385.

Hsu, S.D., Chu, C.H., Tsou, A.P., Chen, S.J., Chen, H.C., Hsu, P.W., Wong, Y.H., Chen, Y.H., Chen, G.H., and Huang, H.D. (2008). miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. Nucleic Acids Res *36*, D165-169.

Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., *et al.* (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res *42*, D78-85.

Hu, X., Schwarz, J.K., Lewis, J.S., Jr., Huettner, P.C., Rader, J.S., Deasy, J.O., Grigsby, P.W., and Wang, X. (2010). A microRNA expression signature for cervical cancer prognosis. Cancer Res *70*, 1441-1448.

Hu, Y., Zhu, Y.Y., Zhang, S.H., Zhu, H., and Shuai, C.X. (2011). Human papillomavirus type 16 e6 gene variations in young Chinese women with cervical squamous cell carcinoma. Reprod Sci *18*, 406-412.

Hu, Z., and Ma, D. (2018). The precision prevention and therapy of HPV-related cervical cancer: new concepts and clinical implications. Cancer Med *7*, 5217-5236.

Hu, Z., Yu, L., Zhu, D., Ding, W., Wang, X., Zhang, C., Wang, L., Jiang, X., Shen, H., He, D., *et al.* (2014). Disruption of HPV16-E7 by CRISPR/Cas system induces apoptosis and growth inhibition in HPV16 positive human cervical cancer cells. Biomed Res Int *2014*, 612823.

Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., Ding, W., Yu, L., Wang, X., Wang, L., *et al.* (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat Genet *47*, 158-163.

Huang, Y., Lau, S.K., Woo, P.C., and Yuen, K.Y. (2008). CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes. Nucleic Acids Res *36*, D504-511.

Hubert, W.G. (2005). Variant upstream regulatory region sequences differentially regulate human papillomavirus type 16 DNA replication throughout the viral life cycle. J Virol *79*, 5914-5922.

Huh, W.K., Joura, E.A., Giuliano, A.R., Iversen, O.E., de Andrade, R.P., Ault, K.A., Bartholomew, D., Cestero, R.M., Fedrizzi, E.N., Hirschberg, A.L., *et al.* (2017). Final efficacy, immunogenicity, and safety analyses of a nine-valent human

papillomavirus vaccine in women aged 16-26 years: a randomised, double-blind trial. Lancet *390*, 2143-2159.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res *39*, D576-582.

Hung, C.F., Ma, B., Monie, A., Tsen, S.W., and Wu, T.C. (2008). Therapeutic human papillomavirus vaccines: current clinical trials and future directions. Expert Opin Biol Ther *8*, 421-439.

Hunt, M., Gall, A., Ong, S.H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J.A., Kellam, P., and Otto, T.D. (2015). IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics *31*, 2374-2376.

Hurwitz, B.L., Ponsero, A., Thornton, J., Jr., and U'Ren, J.M. (2018). Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets. Virus Res *244*, 110-115.

Hussain, M., Taft, R.J., and Asgari, S. (2008). An insect virus-encoded microRNA regulates viral replication. J Virol *82*, 9164-9170.

Hutvagner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056-2060.

Iorio, M.V., and Croce, C.M. (2012). Causes and consequences of microRNA dysregulation. Cancer J *18*, 215-222.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., *et al.* (2005). MicroRNA gene expression deregulation in human breast cancer. Cancer Res *65*, 7065-7070.

Iorio, M.V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., Taccioli, C., Volinia, S., Liu, C.G., Alder, H., *et al.* (2007). MicroRNA signatures in human ovarian cancer. Cancer Res *67*, 8699-8707.

Isakov, O., Bordería, A.V., Golan, D., Hamenahem, A., Celniker, G., Yoffe, L., Blanc, H., Vignuzzi, M., and Shomron, N. (2015). Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. Bioinformatics *31*, 2141-2150.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. Nat Methods *12*, 351-356.

Jansen, K.U., and Shaw, A.R. (2004). Human papillomavirus vaccines and prevention of cervical cancer. Annu Rev Med *55*, 319-331.

Javier, R.T., and Butel, J.S. (2008). The history of tumor virology. Cancer Res *68*, 7693-7706.

Jayakumar, V., and Sakakibara, Y. (2019). Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. Brief Bioinform *20*, 866-876.

Jeon, S., and Lambert, P.F. (1995). Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. Proc Natl Acad Sci U S A *92*, 1654-1658.

Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annu Rev Genet *42*, 709-732.

Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res *45*, W24-W29.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res *37*, D98-104.

Jones, R.E., Wegrzyn, R.J., Patrick, D.R., Balishin, N.L., Vuocolo, G.A., Riemen, M.W., Defeo-Jones, D., Garsky, V.M., Heimbrook, D.C., and Oliff, A. (1990). Identification of HPV-16 E7 peptides that are potent antagonists of E7 binding to the retinoblastoma suppressor protein. J Biol Chem *265*, 12782-12785.

Joura, E.A., Giuliano, A.R., Iversen, O.E., Bouchard, C., Mao, C., Mehlsen, J., Moreira, E.D., Jr., Ngan, Y., Petersen, L.K., Lazcano-Ponce, E., *et al.* (2015). A 9-valent HPV vaccine against infection and intraepithelial neoplasia in women. N Engl J Med *372*, 711-723.

Jung, H.S., Rajasekaran, N., Ju, W., and Shin, Y.K. (2015). Human Papillomavirus: Current and Future RNAi Therapeutic Strategies for Cervical Cancer. J Clin Med *4*, 1126-1155.

Jurtz, V.I., Villarroel, J., Lund, O., Voldby Larsen, M., and Nielsen, M. (2016). MetaPhinder-Identifying Bacteriophage Sequences in Metagenomic Data Sets. PLoS One *11*, e0163111.

Kalantari, M., Chase, D.M., Tewari, K.S., and Bernard, H.U. (2010). Recombination of human papillomavirus-16 and host DNA in exfoliated cervical cells: a pilot study of L1 gene methylation and chromosomal integration as biomarkers of carcinogenic progression. J Med Virol *82*, 311-320.

Kamath, G.M., Shomorony, I., Xia, F., Courtade, T.A., and Tse, D.N. (2017). HINGE: long-read assembly achieves optimal repeat resolution. Genome Res *27*, 747-756.

Kamdar, M.R., and Dumontier, M. (2015). An Ebola virus-centered knowledge base. Database (Oxford) *2015*, bav049.

Kammer, C., Tommasino, M., Syrjanen, S., Delius, H., Hebling, U., Warthorst, U., Pfister, H., and Zehbe, I. (2002). Variants of the long control region and the E6 oncogene in European human papillomavirus type 16 isolates: implications for cervical disease. Br J Cancer *86*, 269-273.

Kammer, C., Warthorst, U., Torrez-Martinez, N., Wheeler, C.M., and Pfister, H. (2000). Sequence analysis of the long control region of human papillomavirus type 16 variants and functional consequences for P97 promoter activity. J Gen Virol *81*, 1975-1981.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res *40*, D109-114.

Kang, S., Jeon, Y.T., Kim, J.W., Park, N.H., Song, Y.S., Kang, S.B., and Lee, H.P. (2005). Polymorphism in the E6 gene of human papillomavirus type 16 in the cervical tissues of Korean women. Int J Gynecol Cancer *15*, 107-112.

Karyala, P., Metri, R., Bathula, C., Yelamanchi, S.K., Sahoo, L., Arjunan, S., Sastri, N.P., and Chandra, N. (2016). DenHunt - A Comprehensive Database of the Intricate Network of Dengue-Human Interactions. PLoS Negl Trop Dis *10*, e0004965.

Kather, A., Ferrara, A., Nonn, M., Schinz, M., Nieland, J., Schneider, A., Durst, M., and Kaufmann, A.M. (2003). Identification of a naturally processed HLA-A*0201 HPV18 E7 T cell epitope by tumor cell mediated in vitro vaccination. Int J Cancer *104*, 345-353.

Katzourakis, A., and Gifford, R.J. (2010). Endogenous viral elements in animal genomes. PLoS Genet *6*, e1001191.

Kaufmann, A.M., Stern, P.L., Rankin, E.M., Sommer, H., Nuessler, V., Schneider, A., Adams, M., Onon, T.S., Bauknecht, T., Wagner, U., *et al.* (2002). Safety and immunogenicity of TA-HPV, a recombinant vaccinia virus expressing modified human papillomavirus (HPV)-16 and HPV-18 E6 and E7 genes, in women with progressive cervical cancer. Clin Cancer Res *8*, 3676-3685.

Kaur, K., Gupta, A.K., Rajput, A., and Kumar, M. (2016). ge-CRISPR - An integrated pipeline for the prediction and analysis of sgRNAs genome editing efficiency for CRISPR/Cas system. Sci Rep *6*, 30870.

Kaya, K.D., Karakulah, G., Yakicier, C.M., Acar, A.C., and Konu, O. (2011). mESAdb: microRNA expression and sequence analysis database. Nucleic Acids Res *39*, D170-180.

Keam, S.J., and Harper, D.M. (2008). Human papillomavirus types 16 and 18 vaccine (recombinant, AS04 adjuvanted, adsorbed) [Cervarix]. Drugs *68*, 359-372.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol *11*, R116.

Kennedy, E.M., Kornepati, A.V., Goldstein, M., Bogerd, H.P., Poling, B.C., Whisnant, A.W., Kastan, M.B., and Cullen, B.R. (2014). Inactivation of the human papillomavirus E6 or E7 gene in cervical carcinoma cells by using a bacterial CRISPR/Cas RNA-guided endonuclease. J Virol *88*, 11965-11972.

Kenter, G.G., Welters, M.J., Valentijn, A.R., Lowik, M.J., Berends-van der Meer, D.M., Vloon, A.P., Essahsah, F., Fathers, L.M., Offringa, R., Drijfhout, J.W., *et al.* (2009). Vaccination against HPV-16 oncoproteins for vulvar intraepithelial neoplasia. N Engl J Med *361*, 1838-1847.

Khan, M.S., Gupta, A.K., and Kumar, M. (2016). ViralEpi v1.0: a high-throughput spectrum of viral epigenomic methylation profiles from diverse diseases. Epigenomics *8*, 67-75.

Kim, K., Garner-Hamrick, P.A., Fisher, C., Lee, D., and Lambert, P.F. (2003). Methylation patterns of papillomavirus DNA, its influence on E2 function, and implications in viral infection. J Virol *77*, 12450-12459.

Kim, M., Zhang, X., Ligo, J.G., Farnoud, F., Veeravalli, V.V., and Milenkovic, O. (2016). MetaCRAM: an integrated pipeline for metagenomic taxonomy identification and compression. BMC Bioinformatics *17*, 94.

Kim, M.K., Kim, H.S., Kim, S.H., Oh, J.M., Han, J.Y., Lim, J.M., Juhnn, Y.S., and Song, Y.S. (2010). Human papillomavirus type 16 E5 oncoprotein as a new target for cervical cancer treatment. Biochem Pharmacol *80*, 1930-1935.

Kim, V.N. (2005a). MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol *6*, 376-385.

Kim, V.N. (2005c). Small RNAs: classification, biogenesis, and function. Mol Cells *19*, 1-15.

Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., Lundegaard, C., Sette, A., Lund, O., Bourne, P.E., *et al.* (2012). Immune epitope database analysis resource. Nucleic Acids Res *40*, W525-530.

Kincaid, R.P., and Sullivan, C.S. (2012). Virus-encoded microRNAs: an overview and a look to the future. PLoS Pathog *8*, e1003018.

Klaes, R., Woerner, S.M., Ridder, R., Wentzensen, N., Duerst, M., Schneider, A., Lotz, B., Melsheimer, P., and von Knebel Doeberitz, M. (1999). Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. Cancer Res *59*, 6132-6136.

Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I., and Bird, A.P. (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Mol Cell *19*, 667-678.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol *37*, 540-546.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res *27*, 722-736.

Kori, M., and Yalcin Arga, K. (2018). Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. PLoS One *13*, e0200717.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res *42*, D68-73.

Kozuka, T., Aoki, Y., Nakagawa, K., Ohtomo, K., Yoshikawa, H., Matsumoto, K., Yoshiike, K., and Kanda, T. (2000). Enhancer-promoter activity of human papillomavirus type 16 long control regions isolated from cell lines SiHa and CaSki and cervical cancer biopsies. Jpn J Cancer Res *91*, 271-279.

Kristensen, D.M., Mushegian, A.R., Dolja, V.V., and Koonin, E.V. (2010). New dimensions of the virus world discovered through metagenomics. Trends Microbiol *18*, 11-19.

Krump, N.A., and You, J. (2018). Molecular mechanisms of viral oncogenesis in humans. Nat Rev Microbiol *16*, 684-698.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.

Kuiken, C., Hraber, P., Thurmond, J., and Yusim, K. (2008). The hepatitis C sequence database in Los Alamos. Nucleic Acids Res *36*, D512-516.

Kuiken, C., Korber, B., and Shafer, R.W. (2003). HIV sequence databases. AIDS Rev *5*, 52-61.

Kuiken, C., Thurmond, J., Dimitrijevic, M., and Yoon, H. (2012). The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. Nucleic Acids Res *40*, D587-592.

Kuiken, C., Yusim, K., Boykin, L., and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. Bioinformatics *21*, 379-384.

Kulkarni-Kale, U., Bhosle, S., Manjari, G.S., and Kolaskar, A.S. (2004). VirGen: a comprehensive viral genome resource. Nucleic Acids Res *32*, D289-292.

Kumar Gupta, A., and Kumar, M. (2015). HPVbase--a knowledgebase of viral integrations, methylation patterns and microRNAs aberrant expression: As potential biomarkers for Human papillomaviruses mediated carcinomas. Sci Rep *5*, 12522.

Kurvinen, K., Yliskoski, M., Saarikoski, S., Syrjanen, K., and Syrjanen, S. (2000). Variants of the long control region of human papillomavirus type 16. Eur J Cancer *36*, 1402-1410.

Küry, P., Nath, A., Créange, A., Dolei, A., Marche, P., Gold, J., Giovannoni, G., Hartung, H.P., and Perron, H. (2018). Human Endogenous Retroviruses in Neurological Diseases. Trends Mol Med *24*, 379-394.

Laehnemann, D., Borkhardt, A., and McHardy, A.C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform *17*, 154-179.

Lai, B., Ding, R., Li, Y., Duan, L., and Zhu, H. (2012). A de novo metagenomic assembly program for shotgun DNA reads. Bioinformatics *28*, 1455-1462.

Lai, B., Wang, F., Wang, X., Duan, L., and Zhu, H. (2015). InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics *16*, 244.

Lajer, C.B., Garnaes, E., Friis-Hansen, L., Norrild, B., Therkildsen, M.H., Glud, M., Rossing, M., Lajer, H., Svane, D., Skotte, L., *et al.* (2012). The role of miRNAs in human papilloma virus (HPV)-associated cancers: bridging between HPV-related head and neck cancer and cervical cancer. Br J Cancer *106*, 1526-1534.

Lambert, C., Braxton, C., Charlebois, R.L., Deyati, A., Duncan, P., La Neve, F., Malicki, H.D., Ribrioux, S., Rozelle, D.K., Michaels, B., *et al.* (2018). Considerations for Optimization of High-Throughput Sequencing Bioinformatics Pipelines for Virus Detection. Viruses *10*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

LaPierre, N., Mangul, S., Alser, M., Mandric, I., Wu, N.C., Koslicki, D., and Eskin, E. (2019). MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. BMC Genomics *20*, 423.

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: de novo assembly for metagenomes. J Comput Biol *18*, 429-443.

Lauss, M., Visne, I., Weinhaeusel, A., Vierlinger, K., Noehammer, C., and Kriegner, A. (2008). MethCancerDB--aberrant DNA methylation in human cancer. Br J Cancer *98*, 816-817.

Lecuit, M., and Eloit, M. (2013). The human virome: new tools and concepts. Trends Microbiol *21*, 510-515.

Lee, J.W., Choi, C.H., Choi, J.J., Park, Y.A., Kim, S.J., Hwang, S.Y., Kim, W.Y., Kim, T.J., Lee, J.H., Kim, B.G., *et al.* (2008a). Altered MicroRNA expression in cervical carcinomas. Clin Cancer Res *14*, 2535-2542.

Lee, K., Magalhaes, I., Clavel, C., Briolat, J., Birembaut, P., Tommasino, M., and Zehbe, I. (2008d). Human papillomavirus 16 E6, L1, L2 and E2 gene variants in cervical lesion progression. Virus Res *131*, 106-110.

Lee, S., Paulson, K.G., Murchison, E.P., Afanasiev, O.K., Alkan, C., Leonard, J.H., Byrd, D.R., Hannon, G.J., and Nghiem, P. (2011). Identification and validation of a novel mature microRNA encoded by the Merkel cell polyomavirus in human Merkel cell carcinomas. J Clin Virol *52*, 272-275.

Leggett, R.M., Heavens, D., Caccamo, M., Clark, M.D., and Davey, R.P. (2016). NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics *32*, 142-144.

Li, B., Hu, Y., Ye, F., Li, Y., Lv, W., and Xie, X. (2010a). Reduced miR-34a expression in normal cervical tissues and cervical lesions with high-risk human papillomavirus infection. Int J Gynecol Cancer *20*, 597-604.

Li, D., Huang, Y., Leung, C.M., Luo, R., Ting, H.F., and Lam, T.W. (2017). MegaGTA: a sensitive and accurate metagenomic gene-targeted assembler using iterative de Bruijn graphs. BMC Bioinformatics *18*, 408.

Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674-1676.

Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., Yamashita, H., and Lam, T.W. (2016a). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3-11.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics *32*, 2103-2110.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, J.W., Wan, R., Yu, C.S., Co, N.N., Wong, N., and Chan, T.F. (2013a). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics *29*, 649-651.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.* (2010b). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res *20*, 265-272.

Li, S.C., Shiau, C.K., and Lin, W.C. (2008). Vir-Mir db: prediction of viral microRNA candidate hairpins. Nucleic Acids Res *36*, D184-189.

Li, W., Zeng, X., Lee, N.P., Liu, X., Chen, S., Guo, B., Yi, S., Zhuang, X., Chen, F., Wang, G., *et al.* (2013b). HIVID: an efficient method to detect HBV integration using low coverage sequencing. Genomics *102*, 338-344.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res *42*, D1070-1074.

Li, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P., and Ma, X. (2016c). VIP: an integrated pipeline for metagenomics of virus identification and discovery. Sci Rep *6*, 23774.

Liao, S., Deng, D., Zhang, W., Hu, X., Wang, W., Wang, H., Lu, Y., Wang, S., Meng, L., and Ma, D. (2013a). Human papillomavirus 16/18 E5 promotes cervical cancer cell proliferation, migration and invasion in vitro and accelerates tumor growth in vivo. Oncol Rep *29*, 95-102.

Liao, S.J., Deng, D.R., Zeng, D., Zhang, L., Hu, X.J., Zhang, W.N., Li, L., Jiang, X.F., Wang, C.Y., Zhou, J.F., *et al.* (2013b). HPV16 E5 peptide vaccine in treatment of cervical cancer in vitro and in vivo. J Huazhong Univ Sci Technolog Med Sci *33*, 735-742.

Lichtig, H., Algrisi, M., Botzer, L.E., Abadi, T., Verbitzky, Y., Jackman, A., Tommasino, M., Zehbe, I., and Sherman, L. (2006). HPV16 E6 natural variants exhibit different activities in functional assays relevant to the carcinogenic potential of E6. Virology *350*, 216-227.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics *24*, 863-865.

Lin, H.H., and Liao, Y.C. (2017). drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. Gigascience *6*, 1-10.

Lin, J., Kramna, L., Autio, R., Hyöty, H., Nykter, M., and Cinek, O. (2017). Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. BMC Genomics *18*, 378.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., and Pevzner, P.A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci U S A *113*, E8396-E8405.

Ling, H., Fabbri, M., and Calin, G.A. (2013). MicroRNAs and other non-coding RNAs as targets for anticancer drug development. Nat Rev Drug Discov *12*, 847-865.

Liu, D.W., Tsao, Y.P., Hsieh, C.H., Hsieh, J.T., Kung, J.T., Chiang, C.L., Huang, S.J., and Chen, S.L. (2000). Induction of CD8 T cells by vaccination with recombinant adenovirus expressing human papillomavirus type 16 E5 gene reduces tumor growth. J Virol *74*, 9083-9089.

Liu, D.W., Yang, Y.C., Lin, H.F., Lin, M.F., Cheng, Y.W., Chu, C.C., Tsao, Y.P., and Chen, S.L. (2007). Cytotoxic T-lymphocyte responses to human papillomavirus type 16 E5 and E7 proteins and HLA-A*0201-restricted T-cell peptides in cervical cancer patients. J Virol *81*, 2869-2879.

Loman, N.J., and Quinlan, A.R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics *30*, 3399-3401.

Londesborough, P., Ho, L., Terry, G., Cuzick, J., Wheeler, C., and Singer, A. (1996). Human papillomavirus genotype as a predictor of persistence and development of high-grade lesions in women with minor cervical abnormalities. Int J Cancer *69*, 364-368.

Longworth, M.S., and Laimins, L.A. (2004). Pathogenesis of human papillomaviruses in differentiating epithelia. Microbiol Mol Biol Rev *68*, 362-372.

López-Cortés, A., Paz, Y.M.C., Guerrero, S., Cabrera-Andrade, A., Barigye, S.J., Munteanu, C.R., González-Díaz, H., Pazos, A., Pérez-Castillo, Y., and Tejera, E. (2020). OncoOmics approaches to reveal essential genes in breast cancer: a panoramic view from pathogenesis to precision medicine. Sci Rep *10*, 5285.

Lorenzi, H.A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., and Williamson, S.J. (2011). TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. Stand Genomic Sci *4*, 418-429.

Lorincz, A.T. (2011). The Promise and the Problems of Epigenetics Biomarkers in Cancer. Expert Opin Med Diagn *5*, 375-379.

Lorincz, A.T., Brentnall, A.R., Vasiljevic, N., Scibior-Bentkowska, D., Castanon, A., Fiander, A., Powell, N., Tristram, A., Cuzick, J., and Sasieni, P. (2013). HPV16 L1 and L2 DNA methylation predicts high-grade cervical intraepithelial neoplasia in women with mildly abnormal cervical cytology. Int J Cancer *133*, 637-644.

Lowy, D.R., and Schiller, J.T. (2006). Prophylactic human papillomavirus vaccines. J Clin Invest *116*, 1167-1173.

Lu, G., Rowley, T., Garten, R., and Donis, R.O. (2007). FluGenome: a web tool for genotyping influenza A virus. Nucleic Acids Res *35*, W275-279.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics *14*, 265-279.

Lui, W.O., Pourmand, N., Patterson, B.K., and Fire, A. (2007). Patterns of known and novel small RNAs in human cervical cancer. Cancer Res *67*, 6031-6043.

Lv, J., Liu, H., Su, J., Wu, X., Li, B., Xiao, X., Wang, F., Wu, Q., and Zhang, Y. (2012). DiseaseMeth: a human disease methylation database. Nucleic Acids Res *40*, D1030-1035.

Maarala, A.I., Bzhalava, Z., Dillner, J., Heljanko, K., and Bzhalava, D. (2018). ViraPipe: scalable parallel pipeline for viral metagenome analysis from next generation sequencing reads. Bioinformatics *34*, 928-935.

Macalalad, A.R., Zody, M.C., Charlebois, P., Lennon, N.J., Newman, R.M., Malboeuf, C.M., Ryan, E.M., Boutwell, C.L., Power, K.A., Brackney, D.E., *et al.* (2012). Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol *8*, e1002417.

Maccallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T.P., *et al.* (2009). ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol *10*, R103.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L.J., and Salzberg, S.L. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics *29*, 1718-1725.

Mangul, S., Wu, N.C., Mancuso, N., Zelikovsky, A., Sun, R., and Eskin, E. (2014). Accurate viral population assembly from ultra-deep sequencing data. Bioinformatics *30*, i329-337.

Manso, C.F., Bibby, D.F., and Mbisa, J.L. (2017). Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. Sci Rep *7*, 4173.

Mantovani, F., and Banks, L. (2001). The human papillomavirus E6 protein and its contribution to malignant progression. Oncogene *20*, 7874-7887.

Manzo-Merino, J., Thomas, M., Fuentes-Gonzalez, A.M., Lizano, M., and Banks, L. (2013). HPV E6 oncoprotein as a potential therapeutic target in HPV related cancers. Expert Opin Ther Targets *17*, 1357-1368.

Martin, D., and Gutkind, J.S. (2008). Human tumor-associated viruses and new insights into the molecular mechanisms of cancer. Oncogene *27 Suppl 2*, S31-42.

Martinez, I., Gardiner, A.S., Board, K.F., Monzon, F.A., Edwards, R.P., and Khan, S.A. (2008). Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells. Oncogene *27*, 2575-2582.

Masson, P., Hulo, C., De Castro, E., Bitter, H., Gruenbaum, L., Essioux, L., Bougueleret, L., Xenarios, I., and Le Mercier, P. (2013). ViralZone: recent updates to the virus knowledge resource. Nucleic Acids Res *41*, D579-583.

Matovina, M., Sabol, I., Grubisic, G., Gasperov, N.M., and Grce, M. (2009). Identification of human papillomavirus type 16 integration sites in high-grade precancerous cervical lesions. Gynecol Oncol *113*, 120-127.

Matsumoto, K., Yasugi, T., Nakagawa, S., Okubo, M., Hirata, R., Maeda, H., Yoshikawa, H., and Taketani, Y. (2003). Human papillomavirus type 16 E6 variants and HLA class II alleles among Japanese women with cervical cancer. Int J Cancer *106*, 919-922.

Matsumoto, K., Yoshikawa, H., Nakagawa, S., Tang, X., Yasugi, T., Kawana, K., Sekiya, S., Hirai, Y., Kukimoto, I., Kanda, T., *et al.* (2000). Enhanced oncogenicity of human papillomavirus type 16 (HPV16) variants in Japanese population. Cancer Lett *156*, 159-165.

Maufort, J.P., Shai, A., Pitot, H.C., and Lambert, P.F. (2010). A role for HPV16 E5 in cervical carcinogenesis. Cancer Res *70*, 2924-2931.

Mazumder Indra, D., Singh, R.K., Mitra, S., Dutta, S., Chakraborty, C., Basu, P.S., Mondal, R.K., Roychoudhury, S., and Panda, C.K. (2011). Genetic and epigenetic changes of HPV16 in cervical cancer differentially regulate E6/E7

expression and associate with disease progression. Gynecol Oncol *123*, 597-604.

McBride, A.A., and Warburton, A. (2017). The role of integration in oncogenic progression of HPV-associated cancers. PLoS Pathog *13*, e1006211.

McNair, K., Zhou, C., Dinsdale, E.A., Souza, B., and Edwards, R.A. (2019). PHANOTATE: a novel approach to gene identification in phage genomes. Bioinformatics *35*, 4537-4542.

ME, J.W., Adair, K., and Brierley, L. (2013). RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases. Microbiol Spectr *1*.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D.I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics *12*, 451.

Megraw, M., Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. (2007). miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res *35*, D149-155.

Meng, Y., Gou, L., Chen, D., Mao, C., Jin, Y., Wu, P., and Chen, M. (2011). PmiRKB: a plant microRNA knowledge base. Nucleic Acids Res *39*, D181-187.

Mesri, E.A., Feitelson, M.A., and Munger, K. (2014). Human viral oncogenesis: a cancer hallmarks analysis. Cell Host Microbe *15*, 266-282.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics *34*, i142-i150.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics *32*, 1088-1090.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315-327.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol *12*, R112.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A *110*, 12450-12455.

Mirabello, L., Schiffman, M., Ghosh, A., Rodriguez, A.C., Vasiljevic, N., Wentzensen, N., Herrero, R., Hildesheim, A., Wacholder, S., Scibior-Bentkowska, D., *et al.* (2013). Elevated methylation of HPV16 DNA is associated with the development of high grade cervical intraepithelial neoplasia. Int J Cancer *132*, 1412-1422.

Mirabello, L., Sun, C., Ghosh, A., Rodriguez, A.C., Schiffman, M., Wentzensen, N., Hildesheim, A., Herrero, R., Wacholder, S., Lorincz, A., *et al.* (2012). Methylation of human papillomavirus type 16 genome and risk of cervical precancer in a Costa Rican population. J Natl Cancer Inst *104*, 556-565.

Mirabello, L., Yeager, M., Yu, K., Clifford, G.M., Xiao, Y., Zhu, B., Cullen, M., Boland, J.F., Wentzensen, N., Nelson, C.W., *et al.* (2017). HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. Cell *170*, 1164-1174 e1166.

Misra, M., and Schein, C.H. (2007). Flavitrack: an annotated database of flavivirus sequences. Bioinformatics *23*, 2645-2647.

Mokili, J.L., Rohwer, F., and Dutilh, B.E. (2012). Metagenomics and future perspectives in virus discovery. Curr Opin Virol *2*, 63-77.

Moody, C.A., and Laimins, L.A. (2010). Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer *10*, 550-560.

Moore, P.S., and Chang, Y. (2010). Why do viruses cause cancer? Highlights of the first century of human tumour virology. Nat Rev Cancer *10*, 878-889.

Morales-Sánchez, A., and Fuentes-Pananá, E.M. (2014). Human viruses and cancer. Viruses *6*, 4047-4079.

Morgan, E.L., and Macdonald, A. (2020). Manipulation of JAK/STAT Signalling by High-Risk HPVs: Potential Therapeutic Targets for HPV-Associated Malignancies. Viruses *12*.

Morishima, S., Akatsuka, Y., Nawa, A., Kondo, E., Kiyono, T., Torikai, H., Nakanishi, T., Ito, Y., Tsujimura, K., Iwata, K., *et al.* (2007). Identification of an HLA-A24-restricted cytotoxic T lymphocyte epitope from human papillomavirus type-16 E6: the combined effects of bortezomib and interferon-gamma on the presentation of a cryptic epitope. Int J Cancer *120*, 594-604.

Morrow, M.P., Yan, J., and Sardesai, N.Y. (2013). Human papillomavirus therapeutic vaccines: targeting viral antigens as immunotherapy for precancerous disease and cancer. Expert Rev Vaccines *12*, 271-283.

Moura, G., Pinheiro, M., Silva, R., Miranda, I., Afreixo, V., Dias, G., Freitas, A., Oliveira, J.L., and Santos, M.A. (2005). Comparative context analysis of codon pairs on an ORFeome scale. Genome Biol *6*, R28.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K.E., Venter, J.C., *et al.* (2017). The blood DNA virome in 8,000 humans. PLoS Pathog *13*, e1006292.

Mui, U.N., Haley, C.T., and Tyring, S.K. (2017). Viral Oncology: Molecular Biology and Pathogenesis. J Clin Med *6*.

Munger, K., Baldwin, A., Edwards, K.M., Hayakawa, H., Nguyen, C.L., Owens, M., Grace, M., and Huh, K. (2004). Mechanisms of human papillomavirus-induced oncogenesis. J Virol *78*, 11451-11460.

Munger, K., Basile, J.R., Duensing, S., Eichten, A., Gonzalez, S.L., Grace, M., and Zacny, V.L. (2001). Biological activities and molecular targets of the human papillomavirus E7 oncoprotein. Oncogene *20*, 7888-7898.

Munoz, N., Bosch, F.X., de Sanjose, S., Herrero, R., Castellsague, X., Shah, K.V., Snijders, P.J., and Meijer, C.J. (2003). Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med *348*, 518-527.

Munoz, N., Castellsague, X., de Gonzalez, A.B., and Gissmann, L. (2006). Chapter 1: HPV in the etiology of human cancer. Vaccine *24 Suppl 3*, S3/1-10.

Murphy, E., Vanicek, J., Robins, H., Shenk, T., and Levine, A.J. (2008). Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: implications for latency. Proc Natl Acad Sci U S A *105*, 5453-5458.

Naccache, S.N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A.L., Luk, K.C., Enge, B., *et al.* (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res *24*, 1180-1192.

Nadiminty, N., Tummala, R., Lou, W., Zhu, Y., Shi, X.B., Zou, J.X., Chen, H., Zhang, J., Chen, X., Luo, J., *et al.* (2012). MicroRNA let-7c is downregulated in prostate cancer and suppresses prostate cancer growth. PLoS One *7*, e32832.

Nagpal, G., Sharma, M., Kumar, S., Chaudhary, K., Gupta, S., Gautam, A., and Raghava, G.P. (2014). PCMdb: pancreatic cancer methylation database. Sci Rep *4*, 4197.

Nakagawa, M., Kim, K.H., and Moscicki, A.B. (2004). Different methods of identifying new antigenic epitopes of human papillomavirus type 16 E6 and E7 proteins. Clin Diagn Lab Immunol *11*, 889-896.

Nakagawa, S., and Takahashi, M.U. (2016). gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. Database (Oxford) *2016*.

Nakamura, Y., Yasuike, M., Nishiki, I., Iwasaki, Y., Fujiwara, A., Kawato, Y., Nakai, T., Nagai, S., Kobayashi, T., Gojobori, T., *et al.* (2016). V-GAP: Viral genome assembly pipeline. Gene *576*, 676-680.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res *40*, e155.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Narayanasamy, S., Jarosz, Y., Muller, E.E., Heintz-Buschart, A., Herold, M., Kaysen, A., Laczny, C.C., Pinel, N., May, P., and Wilmes, P. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol *17*, 260.

Nindl, I., Rindfleisch, K., Lotz, B., Schneider, A., and Durst, M. (1999). Uniform distribution of HPV 16 E6 and E7 variants in patients with normal histology, cervical intra-epithelial neoplasia and cervical cancer. Int J Cancer *82*, 203-207.

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M.P.G. (2018). Overview of Virus Metagenomic Classification Methods and Their Biological Applications. Front Microbiol *9*, 749.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res *27*, 824-834.

Ojesina, A.I., Lichtenstein, L., Freeman, S.S., Pedamallu, C.S., Imaz-Rosshandler, I., Pugh, T.J., Cherniack, A.D., Ambrogio, L., Cibulskis, K., Bertelsen, B., *et al.* (2014). Landscape of genomic alterations in cervical carcinomas. Nature *506*, 371-375.

Olsen, L.R., Zhang, G.L., Reinherz, E.L., and Brusic, V. (2011). FLAVIdB: A data mining system for knowledge discovery in flaviviruses with direct applications in immunology and vaccinology. Immunome Res *7*.

Olusola, P., Banerjee, H.N., Philley, J.V., and Dasgupta, S. (2019). Human Papilloma Virus-Associated Cervical Cancer and Health Disparities. Cells *8*.

Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics *12*, 385.

Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S., and Van Criekinge, W. (2008). PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res *36*, D842-846.

Orton, R.J., Gu, Q., Hughes, J., Maabar, M., Modha, S., Vattipally, S.B., Wilkie, G.S., and Davison, A.J. (2016). Bioinformatics tools for analysing viral genomic data. Rev Sci Tech *35*, 271-285.

Oyervides-Muñoz, M.A., Pérez-Maya, A.A., Rodríguez-Gutiérrez, H.F., Gómez-Macias, G.S., Fajardo-Ramírez, O.R., Treviño, V., Barrera-Saldaña, H.A., and Garza-Rodríguez, M.L. (2018). Understanding the HPV integration and its progression to cervical cancer. Infect Genet Evol *61*, 134-144.

Ozen, M., Creighton, C.J., Ozdemir, M., and Ittmann, M. (2008). Widespread deregulation of microRNA expression in human prostate cancer. Oncogene *27*, 1788-1793.

Paavonen, J., Naud, P., Salmeron, J., Wheeler, C.M., Chow, S.N., Apter, D., Kitchener, H., Castellsague, X., Teixeira, J.C., Skinner, S.R., *et al.* (2009). Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women. Lancet *374*, 301-314.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425-430.

Pallante, P., Visone, R., Ferracin, M., Ferraro, A., Berlingieri, M.T., Troncone, G., Chiappetta, G., Liu, C.G., Santoro, M., Negrini, M., *et al.* (2006). MicroRNA deregulation in human thyroid papillary carcinomas. Endocr Relat Cancer *13*, 497-508.

Pande, S., Jain, N., Prusty, B.K., Bhambhani, S., Gupta, S., Sharma, R., Batra, S., and Das, B.C. (2008). Human papillomavirus type 16 variant analysis of E6, E7, and L1 genes and long control region in biopsy samples from cervical cancer patients in north India. J Clin Microbiol *46*, 1060-1066.

Pang, R.T., Leung, C.O., Ye, T.M., Liu, W., Chiu, P.C., Lam, K.K., Lee, K.F., and Yeung, W.S. (2010). MicroRNA-34a suppresses invasion through downregulation of Notch1 and Jagged1 in cervical carcinoma and choriocarcinoma cells. Carcinogenesis *31*, 1037-1044.

Paolini, F., Curzio, G., Cordeiro, M.N., Massa, S., Mariani, L., Pimpinelli, F., de Freitas, A.C., Franconi, R., and Venuti, A. (2017). HPV 16 E5 oncoprotein is expressed in early stage carcinogenesis and can be a target of immunotherapy. Hum Vaccin Immunother *13*, 291-297.

Parfenov, M., Pedamallu, C.S., Gehlenborg, N., Freeman, S.S., Danilova, L., Bristow, C.A., Lee, S., Hadjipanayis, A.G., Ivanova, E.V., Wilkerson, M.D., *et al.* (2014). Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A *111*, 15544-15549.

Patel, D.A., Rozek, L.S., Colacino, J.A., Van Zomeren-Dohm, A., Ruffin, M.T., Unger, E.R., Dolinoy, D.C., Swan, D.C., Onyekwuluje, J., DeGraffinreid, C.R., *et al.* (2012). Patterns of cellular and HPV 16 methylation as biomarkers for cervical neoplasia. J Virol Methods *184*, 84-92.

Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One *7*, e30619.

Peng, S., Tomson, T.T., Trimble, C., He, L., Hung, C.F., and Wu, T.C. (2006). A combination of DNA vaccines targeting human papillomavirus type 16 E6 and E7 generates potent antitumor effects. Gene Ther *13*, 257-265.

Peng, S., Trimble, C., Wu, L., Pardoll, D., Roden, R., Hung, C.F., and Wu, T.C. (2007). HLA-DQB1*02-restricted HPV-16 E7 peptide-specific CD4+ T-cell immune responses correlate with regression of HPV-16-associated high-grade squamous intraepithelial lesions. Clin Cancer Res *13*, 2479-2487.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics *27*, i94-101.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420-1428.

Pérez Sayáns, M., Chamorro Petronacci, C.M., Lorenzo Pouso, A.I., Padín Iruegas, E., Blanco Carrión, A., Suárez Peñaranda, J.M., and García García, A. (2019). Comprehensive Genomic Review of TCGA Head and Neck Squamous Cell Carcinomas (HNSCC). J Clin Med *8*.

Peter, M., Stransky, N., Couturier, J., Hupé, P., Barillot, E., de Cremoux, P., Cottu, P., Radvanyi, F., and Sastre-Garau, X. (2010). Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. J Pathol *221*, 320-330.

Pfeffer, S., and Voinnet, O. (2006). Viruses, microRNAs and cancer. Oncogene *25*, 6211-6219.

Pfeffer, S., Zavolan, M., Grasser, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., *et al.* (2004). Identification of virus-encoded microRNAs. Science *304*, 734-736.

Picconi, M.A., Alonio, L.V., Sichero, L., Mbayed, V., Villa, L.L., Gronda, J., Campos, R., and Teyssie, A. (2003). Human papillomavirus type-16 variants in Quechua aboriginals from Argentina. J Med Virol *69*, 546-552.

Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., *et al.* (2012). ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res *40*, D593-598.

Pientong, C., Wongwarissara, P., Ekalaksananan, T., Swangphon, P., Kleebkaow, P., Kongyingyoes, B., Siriaunkgul, S., Tungsinmunkong, K., and Suthipintawong, C. (2013). Association of human papillomavirus type 16 long control region mutation and cervical cancer. Virol J *10*, 30.

Pillai, M.R., Hariharan, R., Babu, J.M., Lakshmi, S., Chiplunkar, S.V., Patkar, M., Tongaonkar, H., Dinshaw, K., Jayshree, R.S., Reddy, B.K., *et al.* (2009). Molecular variants of HPV-16 associated with cervical cancer in Indian population. Int J Cancer *125*, 91-103.

Piyathilake, C.J., Macaluso, M., Alvarez, R.D., Chen, M., Badiga, S., Edberg, J.C., Partridge, E.E., and Johanning, G.L. (2011). A higher degree of methylation of the HPV 16 E6 gene is associated with a lower likelihood of being diagnosed with cervical intraepithelial neoplasia. Cancer *117*, 957-963.

Poh, W.T., Xia, E., Chin-Inmanu, K., Wong, L.P., Cheng, A.Y., Malasit, P., Suriyaphol, P., Teo, Y.Y., and Ong, R.T. (2013). Viral quasispecies inference from 454 pyrosequencing. BMC Bioinformatics *14*, 355.

Prosperi, M.C., and Salemi, M. (2012). QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics *28*, 132-133.

Qian, K., Pietila, T., Ronty, M., Michon, F., Frilander, M.J., Ritari, J., Tarkkanen, J., Paulin, L., Auvinen, P., and Auvinen, E. (2013). Identification and validation of human papillomavirus encoded microRNAs. PLoS One *8*, e70202.

Qiu, A.D., Wu, E.Q., Yu, X.H., Jiang, C.L., Jin, Y.H., Wu, Y.G., Chen, Y., Shan, Y.M., Zhang, G.N., Fan, Y., *et al.* (2007). HPV prevalence, E6 sequence variation and physical state of HPV16 isolates from patients with cervical cancer in Sichuan, China. Gynecol Oncol *104*, 77-85.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Quiñones-Mateu, M.E., Avila, S., Reyes-Teran, G., and Martinez, M.A. (2014). Deep sequencing: becoming a critical tool in clinical virology. J Clin Virol *61*, 9-19.

Qureshi, A., Kaur, G., and Kumar, M. (2017). AVCpred: an integrated web server for prediction and design of antiviral compounds. Chem Biol Drug Des *89*, 74-83.

Qureshi, A., Rajput, A., Kaur, G., and Kumar, M. (2018). HIVprotI: an integrated web based platform for prediction and design of HIV proteins inhibitors. J Cheminform *10*, 12.

Qureshi, A., Thakur, N., and Kumar, M. (2013a). HIPdb: a database of experimentally validated HIV inhibiting peptides. PLoS One *8*, e54908.

Qureshi, A., Thakur, N., and Kumar, M. (2013b). VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. J Transl Med *11*, 305.

Qureshi, A., Thakur, N., Monga, I., Thakur, A., and Kumar, M. (2014a). VIRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets. Database (Oxford) *2014*.

Qureshi, A., Thakur, N., Tandon, H., and Kumar, M. (2014d). AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Res *42*, D1147-1153.

Radhakrishna Pillai, M., Sreevidya, S., Pollock, B.H., Jayaprakash, P.G., and Herman, B. (2002). Human papillomavirus type 16 E6 and E7 gene variations in Indian cervical cancer. Gynecol Oncol *87*, 268-273.

Rajput, A., Gupta, A.K., and Kumar, M. (2015). Prediction and analysis of quorum sensing peptides based on sequence features. PLoS One *10*, e0120066.

Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P., and Candela, M. (2016). ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics *17*, 165.

Ranieri, D., Belleudi, F., Magenta, A., and Torrisi, M.R. (2015). HPV16 E5 expression induces switching from FGFR2b to FGFR2c and epithelial-mesenchymal transition. Int J Cancer *137*, 61-72.

Rao, Q., Shen, Q., Zhou, H., Peng, Y., Li, J., and Lin, Z. (2012). Aberrant microRNA expression in human cervical carcinomas. Med Oncol *29*, 1242-1248.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res *47*, W191-w198.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69.

Ren, S., Gaykalova, D.A., Guo, T., Favorov, A.V., Fertig, E.J., Tamayo, P., Callejas-Valera, J.L., Allevato, M., Gilardi, M., Santos, J., *et al.* (2020). HPV E2, E4, E5 drive alternative carcinogenic pathways in HPV positive cancers. Oncogene *39*, 6327-6339.

Reshmi, G., and Pillai, M.R. (2008). Beyond HPV: oncomirs as new players in cervical cancer. FEBS Lett *582*, 4113-4116.

Ressing, M.E., Sette, A., Brandt, R.M., Ruppert, J., Wentworth, P.A., Hartman, M., Oseroff, C., Grey, H.M., Melief, C.J., and Kast, W.M. (1995). Human CTL epitopes encoded by human papillomavirus type 16 E6 and E7 identified through in vivo and in vitro immunogenicity studies of HLA-A*0201-binding peptides. J Immunol *154*, 5934-5943.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334-338.

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., Lichter, P., *et al.* (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One *8*, e66621.

Riemer, A.B., Keskin, D.B., Zhang, G., Handley, M., Anderson, K.S., Brusic, V., Reinhold, B., and Reinherz, E.L. (2010). A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers. J Biol Chem *285*, 29608-29622.

Ristriani, T., Fournane, S., Orfanoudakis, G., Trave, G., and Masson, M. (2009). A single-codon mutation converts HPV16 E6 oncoprotein into a potential tumor suppressor, which induces p53-dependent senescence of HPV-positive HeLa cervical cancer cells. Oncogene *28*, 762-772.

Robasky, K., Lewis, N.E., and Church, G.M. (2014). The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet *15*, 56-62.

Robertson, K.D. (2005). DNA methylation and human disease. Nat Rev Genet *6*, 597-610.

Roden, R.B., Ling, M., and Wu, T.C. (2004). Vaccination to prevent and treat cervical cancer. Hum Pathol *35*, 971-982.

Rohwer, F. (2003). Global phage diversity. Cell *113*, 141.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., and Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. Virus Evol *2*, vew022.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol *14*, R51.

Rountree, M.R., Bachman, K.E., Herman, J.G., and Baylin, S.B. (2001). DNA methylation, chromatin inheritance, and cancer. Oncogene *20*, 3156-3165.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., and Enault, F. (2011). Metavir: a web server dedicated to virome analysis. Bioinformatics *27*, 3074-3075.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics *15*, 76.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat Methods *17*, 155-158.

Ruby, J.G., Bellare, P., and Derisi, J.L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) *3*, 865-880.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Rudolf, M.P., Man, S., Melief, C.J., Sette, A., and Kast, W.M. (2001). Human T-cell responses to HLA-A-restricted high binding affinity peptides of human papillomavirus type 18 proteins E6 and E7. Clin Cancer Res *7*, 788s-795s.

Rusan, M., Li, Y.Y., and Hammerman, P.S. (2015). Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res *21*, 2009-2019.

Sahasrabuddhe, V.V., Luhn, P., and Wentzensen, N. (2011). Human papillomavirus and cervical cancer: biomarkers for improved prevention efforts. Future Microbiol *6*, 1083-1098.

Sahli, M., and Shibuya, T. (2012). Arapan-S: a fast and highly accurate whole-genome assembly software for viruses and small genomes. BMC Res Notes *5*, 243.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., *et al.* (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res *22*, 557-567.

Sarkar, A.K., Tortolero-Luna, G., Follen, M., and Sastry, K.J. (2005). Inverse correlation of cellular immune responses specific to synthetic peptides from the E6 and E7 oncoproteins of HPV-16 with recurrence of cervical intraepithelial neoplasia in a cross-sectional study. Gynecol Oncol *99*, S251-261.

Sauvage, V., and Eloit, M. (2016). Viral metagenomics and blood safety. Transfus Clin Biol *23*, 28-38.

Sauvage, V., Laperche, S., Cheval, J., Muth, E., Dubois, M., Boizeau, L., Hébert, C., Lionnet, F., Lefrère, J.J., and Eloit, M. (2016). Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. Blood Transfus *14*, 400-407.

Scarpellini, E., Ianiro, G., Attili, F., Bassanelli, C., De Santis, A., and Gasbarrini, A. (2015). The human gut microbiota and virome: Potential therapeutic implications. Dig Liver Dis *47*, 1007-1012.

Scheffner, M., Werness, B.A., Huibregtse, J.M., Levine, A.J., and Howley, P.M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell *63*, 1129-1136.

Schelhorn, S.E., Fischer, M., Tolosi, L., Altmüller, J., Nürnberg, P., Pfister, H., Lengauer, T., and Berthold, F. (2013). Sensitive detection of viral transcripts in human tumor transcriptomes. PLoS Comput Biol *9*, e1003228.

Schiffman, M., Castle, P.E., Jeronimo, J., Rodriguez, A.C., and Wacholder, S. (2007). Human papillomavirus and cervical cancer. Lancet *370*, 890-907.

Schiffman, M., and Wentzensen, N. (2013). Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. Cancer Epidemiol Biomarkers Prev *22*, 553-560.

Schiffman, M., Wentzensen, N., Wacholder, S., Kinney, W., Gage, J.C., and Castle, P.E. (2011). Human papillomavirus testing in the prevention of cervical cancer. J Natl Cancer Inst *103*, 368-383.

Schmidt, M., Kedzia, W., and Gozdzicka-Jozefiak, A. (2001). Intratype HPV16 sequence variation within LCR of isolates from asymptomatic carriers and cervical cancers. J Clin Virol *23*, 65-77.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863-864.

Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I.B., and Durst, M. (2012). Non-random integration of the HPV genome in cervical cancer. PLoS One *7*, e39632.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.

Schwarz, E., Freese, U.K., Gissmann, L., Mayer, W., Roggenbuck, B., Stremlau, A., and zur Hausen, H. (1985). Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. Nature *314*, 111-114.

Seiwert, T.Y., Zuo, Z., Keck, M.K., Khattri, A., Pedamallu, C.S., Stricker, T., Brown, C., Pugh, T.J., Stojanov, P., Cho, J., *et al.* (2015). Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. Clin Cancer Res *21*, 632-641.

Senol Cali, D., Kim, J.S., Ghose, S., Alkan, C., and Mutlu, O. (2019). Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Brief Bioinform *20*, 1542-1559.

Seo, G.J., Chen, C.J., and Sullivan, C.S. (2009). Merkel cell polyomavirus encodes a microRNA with the ability to autoregulate viral gene expression. Virology *383*, 183-187.

Seo, G.J., Fink, L.H., O'Hara, B., Atwood, W.J., and Sullivan, C.S. (2008). Evolutionarily conserved function of a viral microRNA. J Virol *82*, 9823-9828.

Shang, Q., Wang, Y., Fang, Y., Wei, L., Chen, S., Sun, Y., Li, B., Zhang, F., and Gu, H. (2011). Human papillomavirus type 16 variant analysis of E6, E7, and L1 [corrected] genes and long control region in [corrected] cervical carcinomas in patients in northeast China. J Clin Microbiol *49*, 2656-2663.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shen, S.N., Wang, L.F., Jia, Y.F., Hao, Y.Q., Zhang, L., and Wang, H. (2013). Upregulation of microRNA-224 is associated with aggressive progression and poor prognosis in human cervical cancer. Diagn Pathol *8*, 69.

Shepard, S.S., Meno, S., Bahl, J., Wilson, M.M., Barnes, J., and Neuhaus, E. (2016). Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics *17*, 708.

Shi, M., Zhang, Y.Z., and Holmes, E.C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. Virus Res *243*, 83-90.

Sichero, L., Ferreira, S., Trottier, H., Duarte-Franco, E., Ferenczy, A., Franco, E.L., and Villa, L.L. (2007). High grade cervical lesions are caused preferentially by non-European variants of HPVs 16 and 18. Int J Cancer *120*, 1763-1768.

Siddiqui, M.A., and Perry, C.M. (2006). Human papillomavirus quadrivalent (types 6, 11, 16, 18) recombinant vaccine (Gardasil). Drugs *66*, 1263-1271; discussion 1272-1263.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res *19*, 1117-1123.

Singh, H., Ansari, H.R., and Raghava, G.P. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One *8*, e62216.

Skalsky, R.L., and Cullen, B.R. (2010). Viruses, microRNAs, and host interactions. Annu Rev Microbiol *64*, 123-141.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., and DeRisi, J.L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One *9*, e105067.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J., and Holmes, I.H. (2009). JBrowse: a next-generation genome browser. Genome Res *19*, 1630-1638.

Smith, B., Chen, Z., Reimers, L., van Doorslaer, K., Schiffman, M., Desalle, R., Herrero, R., Yu, K., Wacholder, S., Wang, T., *et al.* (2011). Sequence

imputation of HPV16 genomes for genetic association studies. PLoS One *6*, e21375.

Smith, J.S., Lindsay, L., Hoots, B., Keys, J., Franceschi, S., Winer, R., and Clifford, G.M. (2007). Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. Int J Cancer *121*, 621-632.

Smith, K.L., Tristram, A., Gallagher, K.M., Fiander, A.N., and Man, S. (2005). Epitope specificity and longevity of a vaccine-induced human T cell response against HPV18. Int Immunol *17*, 167-176.

Soeda, E., Ferran, M.C., Baker, C.C., and McBride, A.A. (2006). Repression of HPV16 early region transcription by the E2 protein. Virology *351*, 29-41.

Sommer, D.D., Delcher, A.L., Salzberg, S.L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. BMC Bioinformatics *8*, 64.

Song, W., Sun, H.X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., *et al.* (2019). Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Res *47*, W74-W80.

Song, Y.S., Kee, S.H., Kim, J.W., Park, N.H., Kang, S.B., Chang, W.H., and Lee, H.P. (1997). Major sequence variants in E7 gene of human papillomavirus type 16 from cervical cancerous and noncancerous lesions of Korean women. Gynecol Oncol *66*, 275-281.

Sović, I., Šikić, M., Wilm, A., Fenlon, S.N., Chen, S., and Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun *7*, 11307.

Stanley, M. (2006). Immune responses to human papillomavirus. Vaccine *24 Suppl 1*, S16-22.

Stano, M., Beke, G., and Klucar, L. (2016). viruSITE-integrated database for viral genomics. Database (Oxford) *2016*.

Steller, M.A., Gurski, K.J., Murakami, M., Daniel, R.W., Shah, K.V., Celis, E., Sette, A., Trimble, E.L., Park, R.C., and Marincola, F.M. (1998). Cell-mediated immunological responses in cervical and vaginal cancer patients immunized with a lipidated epitope of human papillomavirus type 16 E7. Clin Cancer Res *4*, 2103-2109.

Stephen, A.L., Thompson, C.H., Tattersall, M.H., Cossart, Y.E., and Rose, B.R. (2000). Analysis of mutations in the URR and E6/E7 oncogenes of HPV 16 cervical cancer isolates from central China. Int J Cancer *86*, 695-701.

Stoppler, M.C., Ching, K., Stoppler, H., Clancy, K., Schlegel, R., and Icenogle, J. (1996). Natural variants of the human papillomavirus type 16 E6 protein differ in their abilities to alter keratinocyte differentiation and to induce p53 degradation. J Virol *70*, 6987-6993.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques *28*, 1102, 1104.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., *et al.* (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157-1160.

Stunkel, W., and Bernard, H.U. (1999). The chromatin structure of the long control region of human papillomavirus type 16 represses viral oncoprotein expression. J Virol *73*, 1918-1930.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sullivan, C.S., Grundhoff, A.T., Tevethia, S., Pipas, J.M., and Ganem, D. (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. Nature *435*, 682-686.

Sun, C., Reimers, L.L., and Burk, R.D. (2011a). Methylation of HPV16 genome CpG sites is associated with cervix precancer and cancer. Gynecol Oncol *121*, 59-63.

Sun, Z., Lu, Z., Liu, J., Wang, G., Zhou, W., Yang, L., Liu, C., Wang, B., and Ruan, Q. (2013). Genetic variations of E6 and long control region of human papillomavirus type 16 from patients with cervical lesion in Liaoning, China. BMC Cancer *13*, 459.

Sun, Z., Ren, G., Cui, X., Zhou, W., Liu, C., and Ruan, Q. (2011b). Genetic diversity of HPV-16 E6, E7, and L1 genes in women with cervical lesions in Liaoning Province, China. Int J Gynecol Cancer *21*, 551-558.

Szczesniak, M.W., and Makalowska, I. (2014). miRNEST 2.0: a database of plant and animal microRNAs. Nucleic Acids Res *42*, D74-77.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res *47*, D607-d613.

Tan, S.H., Leong, L.E., Walker, P.A., and Bernard, H.U. (1994). The human papillomavirus type 16 E2 transcription factor binds with low cooperativity to two flanking sites and represses the E6 promoter through displacement of Sp1 and TFIID. J Virol *68*, 6411-6420.

Tang, T., Wong, H.K., Gu, W., Yu, M.Y., To, K.F., Wang, C.C., Wong, Y.F., Cheung, T.H., Chung, T.K., and Choy, K.W. (2013). MicroRNA-182 plays an onco-miRNA role in cervical cancer. Gynecol Oncol *129*, 199-208.

Tarraga, J., Gallego, A., Arnau, V., Medina, I., and Dopazo, J. (2016). HPG pore: an efficient and scalable framework for nanopore sequencing data. BMC Bioinformatics *17*, 107.

Tennakoon, C., and Sung, W.K. (2017). BATVI: Fast, sensitive and accurate detection of virus integrations. BMC Bioinformatics *18*, 71.

Testa, J.S., and Philip, R. (2012). Role of T-cell epitope-based vaccine in prophylactic and therapeutic applications. Future Virol *7*, 1077-1088.

Thakur, A., Qureshi, A., and Kumar, M. (2017). vhfRNAi: a web-platform for analysis of host genes involved in viral infections discovered by genome wide RNAi screens. Mol Biosyst *13*, 1377-1387.

Thakur, A., Rajput, A., and Kumar, M. (2016). MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. Mol Biosyst *12*, 2572-2586.

Thakur, N., Qureshi, A., and Kumar, M. (2012a). AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res *40*, W199-204.

Thakur, N., Qureshi, A., and Kumar, M. (2012c). VIRsiRNAdb: a curated database of experimentally validated viral siRNA/shRNA. Nucleic Acids Res *40*, D230-236.

Thorland, E.C., Myers, S.L., Gostout, B.S., and Smith, D.I. (2003). Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. Oncogene *22*, 1225-1237.

Tithi, S.S., Aylward, F.O., Jensen, R.V., and Zhang, L. (2018). FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. PeerJ *6*, e4227.

Tjalma, W.A., Arbyn, M., Paavonen, J., van Waes, T.R., and Bogers, J.J. (2004). Prophylactic human papillomavirus vaccines: the beginning of the end of cervical cancer. Int J Gynecol Cancer *14*, 751-761.

Tornesello, M.L., Buonaguro, L., Giorgi-Rossi, P., and Buonaguro, F.M. (2013). Viral and cellular biomarkers in the diagnosis of cervical intraepithelial neoplasia and cancer. Biomed Res Int *2013*, 519619.

Tornesello, M.L., Duraturo, M.L., Salatiello, I., Buonaguro, L., Losito, S., Botti, G., Stellato, G., Greggi, S., Piccoli, R., Pilotti, S., *et al.* (2004). Analysis of human papillomavirus type-16 variants in Italian women with cervical intraepithelial neoplasia and cervical cancer. J Med Virol *74*, 117-126.

Trapnell, C., and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nat Biotechnol *27*, 455-457.

Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A.E., Phillippy, A.M., and Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol *14*, R2.

Tuna, M., and Amos, C.I. (2017). Next generation sequencing and its applications in HPV-associated cancers. Oncotarget *8*, 8877-8889.

Turan, T., Kalantari, M., Calleja-Macias, I.E., Cubie, H.A., Cuschieri, K., Villa, L.L., Skomedal, H., Barrera-Saldana, H.A., and Bernard, H.U. (2006). Methylation of the human papillomavirus-18 L1 gene: a biomarker of neoplastic progression? Virology *349*, 175-183.

Turan, T., Kalantari, M., Cuschieri, K., Cubie, H.A., Skomedal, H., and Bernard, H.U. (2007). High-throughput detection of human papillomavirus-18 L1 gene methylation, a candidate biomarker for the progression of cervical neoplasia. Virology *361*, 185-193.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods *13*, 966-967.

Tyagi, A., Ahmed, F., Thakur, N., Sharma, A., Raghava, G.P., and Kumar, M. (2011). HIVsirDB: a database of HIV inhibiting siRNAs. PLoS One *6*, e25917.

Ueda, M.T., Kryukov, K., Mitsuhashi, S., Mitsuhashi, H., Imanishi, T., and Nakagawa, S. (2020). Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. Mob DNA *11*, 29.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev *20*, 515-524.

Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y., and McBride, A.A. (2017). The Papillomavirus Episteme:

a major update to the papillomavirus sequence database. Nucleic Acids Res *45*, D499-D506.

Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., Huyen, Y., and McBride, A.A. (2013). The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res *41*, D571-578.

van Duin, M., Snijders, P.J., Vossen, M.T., Klaassen, E., Voorhorst, F., Verheijen, R.H., Helmerhorst, T.J., Meijer, C.J., and Walboomers, J.M. (2000). Analysis of human papillomavirus type 16 E6 variants in relation to p53 codon 72 polymorphism genotypes in cervical carcinogenesis. J Gen Virol *81*, 317-325.

Vande Pol, S.B., and Klingelhutz, A.J. (2013). Papillomavirus E6 oncoproteins. Virology *445*, 115-137.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res *27*, 737-746.

Vasiljevic, N., Scibior-Bentkowska, D., Brentnall, A., Cuzick, J., and Lorincz, A. (2014). A comparison of methylation levels in HPV18, HPV31 and HPV33 genomes reveals similar associations with cervical precancers. J Clin Virol *59*, 161-166.

Verbist, B., Clement, L., Reumers, J., Thys, K., Vapirev, A., Talloen, W., Wetzels, Y., Meys, J., Aerssens, J., Bijnens, L., *et al.* (2015a). ViVaMBC: estimating viral sequence variation in complex populations from illumina deep-sequencing data using model-based clustering. BMC Bioinformatics *16*, 59.

Verbist, B.M., Thys, K., Reumers, J., Wetzels, Y., Van der Borght, K., Talloen, W., Aerssens, J., Clement, L., and Thas, O. (2015c). VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. Bioinformatics *31*, 94-101.

Veress, G., Murvai, M., Szarka, K., Juhasz, A., Konya, J., and Gergely, L. (2001). Transcriptional activity of human papillomavirus type 16 variants having deletions in the long control region. Eur J Cancer *37*, 1946-1952.

Veress, G., Szarka, K., Dong, X.P., Gergely, L., and Pfister, H. (1999). Functional significance of sequence variation in the E2 gene and the long control region of human papillomavirus type 16. J Gen Virol *80 (Pt 4)*, 1035-1043.

Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A.G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res *40*, D222-229.

Verneau, J., Levasseur, A., Raoult, D., La Scola, B., and Colson, P. (2016). MG-Digger: An Automated Pipeline to Search for Giant Virus-Related Sequences in Metagenomes. Front Microbiol *7*, 428.

Villa, L.L., Costa, R.L., Petta, C.A., Andrade, R.P., Ault, K.A., Giuliano, A.R., Wheeler, C.M., Koutsky, L.A., Malm, C., Lehtinen, M., *et al.* (2005). Prophylactic quadrivalent human papillomavirus (types 6, 11, 16, and 18) L1 virus-like particle vaccine in young women: a randomised double-blind placebo-controlled multicentre phase II efficacy trial. Lancet Oncol *6*, 271-278.

Villa, L.L., Sichero, L., Rahal, P., Caballero, O., Ferenczy, A., Rohan, T., and Franco, E.L. (2000). Molecular variants of human papillomavirus types 16 and 18 preferentially associated with cervical neoplasia. J Gen Virol *81*, 2959-2968.

Virgin, H.W. (2014). The virome in mammalian physiology and disease. Cell *157*, 142-150.

Visone, R., Pallante, P., Vecchione, A., Cirombella, R., Ferracin, M., Ferraro, A., Volinia, S., Coluzzi, S., Leone, V., Borbone, E., *et al.* (2007). Specific microRNAs are downregulated in human thyroid anaplastic carcinomas. Oncogene *26*, 7590-7595.

Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A., *et al.* (2015). The immune epitope database (IEDB) 3.0. Nucleic Acids Res *43*, D405-412.

Vollmers, J., Wiegand, S., and Kaster, A.K. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS One *12*, e0169662.

Wajid, B., and Serpedin, E. (2012). Review of general algorithmic features for genome assemblers for next generation sequencers. Genomics Proteomics Bioinformatics *10*, 58-73.

Walboomers, J.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J., Peto, J., Meijer, C.J., and Munoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol *189*, 12-19.

Wald, A.I., Hoskins, E.E., Wells, S.I., Ferris, R.L., and Khan, S.A. (2011). Alteration of microRNA profiles in squamous cell carcinoma of the head and neck cell lines by human papillomavirus. Head Neck *33*, 504-512.

Walve, R., Rastas, P., and Salmela, L. (2019). Kermit: linkage map guided long read assembly. Algorithms Mol Biol *14*, 8.

Wan, Y., Renner, D.W., Albert, I., and Szpara, M.L. (2015). VirAmp: a galaxy-based viral genome assembly pipeline. Gigascience *4*, 19.

Wang, J., Lu, M., Qiu, C., and Cui, Q. (2010). TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res *38*, D119-122.

Wang, Q., Fish, J.A., Gilman, M., Sun, Y., Brown, C.T., Tiedje, J.M., and Cole, J.R. (2015a). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome *3*, 32.

Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One *8*, e64465.

Wang, Q., Jia, P., and Zhao, Z. (2015b). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med *7*, 2.

Wang, X., Tang, S., Le, S.Y., Lu, R., Rader, J.S., Meyers, C., and Zheng, Z.M. (2008). Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth. PLoS One *3*, e2557.

Wang, X., Wang, H.K., Li, Y., Hafner, M., Banerjee, N.S., Tang, S., Briskin, D., Meyers, C., Chow, L.T., Xie, X., *et al.* (2014). microRNAs are biomarkers of oncogenic human papillomavirus infections. Proc Natl Acad Sci U S A *111*, 4262-4267.

Wang, X., Wang, H.K., McCoy, J.P., Banerjee, N.S., Rader, J.S., Broker, T.R., Meyers, C., Chow, L.T., and Zheng, Z.M. (2009). Oncogenic HPV infection interrupts the expression of tumor-suppressive miR-34a through viral oncoprotein E6. RNA *15*, 637-647.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet *39*, 380-385.

Warren, R.L., Sutton, G.G., Jones, S.J., and Holt, R.A. (2007). Assembling millions of short DNA sequences using SSAKE. Bioinformatics *23*, 500-501.

Watson, M., Thomson, M., Risse, J., Talbot, R., Santoyo-Lopez, J., Gharbi, K., and Blaxter, M. (2015). poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics *31*, 114-115.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

Wentzensen, N., Sun, C., Ghosh, A., Kinney, W., Mirabello, L., Wacholder, S., Shaber, R., LaMere, B., Clarke, M., Lorincz, A.T., *et al.* (2012). Methylation of HPV18, HPV31, and HPV45 genomes and cervical intraepithelial neoplasia grade 3. J Natl Cancer Inst *104*, 1738-1749.

Wentzensen, N., Vinokurova, S., and von Knebel Doeberitz, M. (2004). Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. Cancer Res *64*, 3878-3884.

Werness, B.A., Levine, A.J., and Howley, P.M. (1990). Association of human papillomavirus types 16 and 18 E6 proteins with p53. Science *248*, 76-79.

Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol *13*, e1005595.

Williams, V.M., Filippova, M., Soto, U., and Duerksen-Hughes, P.J. (2011). HPV-DNA integration and carcinogenesis: putative roles for inflammation and oxidative stress. Future Virol *6*, 45-57.

Wilting, S.M., Snijders, P.J., Verlaat, W., Jaspers, A., van de Wiel, M.A., van Wieringen, W.N., Meijer, G.A., Kenter, G.G., Yi, Y., le Sage, C., *et al.* (2013). Altered microRNA expression associated with chromosomal changes contributes to cervical carcinogenesis. Oncogene *32*, 106-116.

Wilting, S.M., van Boerdonk, R.A., Henken, F.E., Meijer, C.J., Diosdado, B., Meijer, G.A., le Sage, C., Agami, R., Snijders, P.J., and Steenbergen, R.D. (2010). Methylation-mediated silencing and tumour suppressive function of hsa-miR-124 in cervical cancer. Mol Cancer *9*, 167.

Winter, J., and Diederichs, S. (2011). MicroRNA biogenesis and cancer. Methods Mol Biol *676*, 3-22.

Winter, J., Jung, S., Keller, S., Gregory, R.I., and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol *11*, 228-234.

Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J.H., Krupovic, M., Dolja, V.V., and Koonin, E.V. (2018). Origins and Evolution of the Global RNA Virome. mBio *9*.

Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., and Nasko, D.J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci *6*, 427-439.

Woodman, C.B., Collins, S.I., and Young, L.S. (2007). The natural history of cervical HPV infection: unresolved issues. Nat Rev Cancer *7*, 11-22.

Woolhouse, M., Scott, F., Hudson, Z., Howey, R., and Chase-Topping, M. (2012). Human viruses: discovery and emergence. Philos Trans R Soc Lond B Biol Sci *367*, 2864-2871.

Wu, Y., Chen, Y., Li, L., Yu, G., He, Y., and Zhang, Y. (2006). Analysis of mutations in the E6/E7 oncogenes and L1 gene of human papillomavirus 16 cervical cancer isolates from China. J Gen Virol *87*, 1181-1188.

Wylie, K.M., Mihindukulasuriya, K.A., Zhou, Y., Sodergren, E., Storch, G.A., and Weinstock, G.M. (2014). Metagenomic analysis of double-stranded DNA viruses in healthy adults. BMC Biol *12*, 71.

Wylie, K.M., Weinstock, G.M., and Storch, G.A. (2012). Emerging view of the human virome. Transl Res *160*, 283-290.

Wylie, K.M., Weinstock, G.M., and Storch, G.A. (2013). Virome genomics: a tool for defining the human virome. Curr Opin Microbiol *16*, 479-484.

Xi, J., Chen, J., Xu, M., Yang, H., Luo, J., Pan, Y., Wang, X., Qiu, L., Yang, J., and Sun, Q. (2017). Genetic variability and functional implication of the long control region in HPV-16 variants in Southwest China. PLoS One *12*, e0182388.

Xi, L.F., Jiang, M., Shen, Z., Hulbert, A., Zhou, X.H., Lin, Y.Y., Kiviat, N.B., and Koutsky, L.A. (2011). Inverse association between methylation of human papillomavirus type 16 DNA and risk of cervical intraepithelial neoplasia grades 2 or 3. PLoS One *6*, e23897.

Xi, L.F., Koutsky, L.A., Galloway, D.A., Kuypers, J., Hughes, J.P., Wheeler, C.M., Holmes, K.K., and Kiviat, N.B. (1997). Genomic variation of human papillomavirus type 16 and risk for high grade cervical intraepithelial neoplasia. J Natl Cancer Inst *89*, 796-802.

Xi, L.F., Koutsky, L.A., Hildesheim, A., Galloway, D.A., Wheeler, C.M., Winer, R.L., Ho, J., and Kiviat, N.B. (2007). Risk for high-grade cervical intraepithelial neoplasia associated with variants of human papillomavirus types 16 and 18. Cancer Epidemiol Biomarkers Prev *16*, 4-10.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res *37*, D105-110.

Xie, H., Zhao, Y., Caramuta, S., Larsson, C., and Lui, W.O. (2012). miR-205 expression promotes cell proliferation and migration of human cervical cancer cells. PLoS One *7*, e46990.

Xin, Y., Chanrion, B., O'Donnell, A.H., Milekic, M., Costa, R., Ge, Y., and Haghighi, F.G. (2012). MethylomeDB: a database of DNA methylation profiles of the brain. Nucleic Acids Res *40*, D1245-1249.

Xiong, G.W., Yuan, Y., Li, M., Guo, H.Y., and Zhang, X.W. (2010). [Human papillomavirus type 16 variant analysis of upstream regulatory region and E6, E7 oncogene from cervical cancer patients in Beijing]. Yi Chuan *32*, 339-347.

Xu, N., Segerman, B., Zhou, X., and Akusjarvi, G. (2007). Adenovirus virus-associated RNAII-derived small RNAs are efficiently incorporated into the rna-induced silencing complex and associate with polyribosomes. J Virol *81*, 10540-10549.

Xue, B., Sechi, L.A., and Kelvin, D.J. (2020). Human Endogenous Retrovirus K (HML-2) in Health and Disease. Front Microbiol *11*, 1690.

Xue, Y., Bellanger, S., Zhang, W., Lim, D., Low, J., Lunny, D., and Thierry, F. (2010). HPV16 E2 is an immediate early marker of viral infection, preceding E7

expression in precursor structures of cervical carcinoma. Cancer Res *70*, 5316-5325.

Yamada, T., Manos, M.M., Peto, J., Greer, C.E., Munoz, N., Bosch, F.X., and Wheeler, C.M. (1997). Human papillomavirus type 16 sequence variation in cervical cancers: a worldwide perspective. J Virol *71*, 2463-2472.

Yamamoto, N., Kinoshita, T., Nohata, N., Itesako, T., Yoshino, H., Enokida, H., Nakagawa, M., Shozu, M., and Seki, N. (2013). Tumor suppressive microRNA-218 inhibits cancer cell migration and invasion by targeting focal adhesion pathways in cervical squamous cell carcinoma. Int J Oncol *42*, 1523-1532.

Yamashita, A., Sakamoto, T., Sekizuka, T., Kato, K., Takasaki, T., and Kuroda, M. (2016a). DGV: Dengue Genographic Viewer. Front Microbiol *7*, 875.

Yamashita, A., Sekizuka, T., and Kuroda, M. (2016b). VirusTAP: Viral Genome-Targeted Assembly Pipeline. Front Microbiol *7*, 32.

Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M., Zody, M.C., and Henn, M.R. (2012). De novo assembly of highly diverse viral populations. BMC Genomics *13*, 475.

Yang, X., Charlebois, P., Macalalad, A., Henn, M.R., and Zody, M.C. (2013). V-Phaser 2: variant inference for viral populations. BMC Genomics *14*, 674.

Yang, X., Li, M., Liu, Q., Zhang, Y., Qian, J., Wan, X., Wang, A., Zhang, H., Zhu, C., Lu, X., *et al.* (2015). Dr.VIS v2.0: an updated database of human disease-related viral integration sites in the era of high-throughput deep sequencing. Nucleic Acids Res *43*, D887-892.

Yang, Z., Chen, S., Luan, X., Li, Y., Liu, M., Li, X., Liu, T., and Tang, H. (2009). MicroRNA-214 is aberrantly expressed in cervical cancers and inhibits the growth of HeLa cells. IUBMB Life *61*, 1075-1082.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev *17*, 3011-3016.

Yim, E.K., and Park, J.S. (2005). The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. Cancer Res Treat *37*, 319-324.

Youde, S.J., Dunbar, P.R., Evans, E.M., Fiander, A.N., Borysiewicz, L.K., Cerundolo, V., and Man, S. (2000). Use of fluorogenic histocompatibility leukocyte antigen-A*0201/HPV 16 E7 peptide complexes to isolate rare human cytotoxic T-lymphocyte-recognizing endogenous human papillomavirus antigens. Cancer Res *60*, 365-371.

Yunis, J.J., and Soreng, A.L. (1984). Constitutive fragile sites and cancer. Science *226*, 1199-1204.

Zaman, M.S., Chen, Y., Deng, G., Shahryari, V., Suh, S.O., Saini, S., Majid, S., Liu, J., Khatri, G., Tanaka, Y., *et al.* (2010). The functional significance of microRNA-145 in prostate cancer. Br J Cancer *103*, 256-264.

Zandberg, D.P., Bhargava, R., Badin, S., and Cullen, K.J. (2013). The role of human papillomavirus in nongenital cancers. CA Cancer J Clin *63*, 57-81.

Zehbe, I., Tachezy, R., Mytilineos, J., Voglino, G., Mikyskova, I., Delius, H., Marongiu, A., Gissmann, L., Wilander, E., and Tommasino, M. (2001). Human papillomavirus 16 E6 polymorphisms in cervical lesions from different European populations and their correlation with human leukocyte antigen class II haplotypes. Int J Cancer *94*, 711-716.

Zehbe, I., Wilander, E., Delius, H., and Tommasino, M. (1998). Human papillomavirus 16 E6 variants are more prevalent in invasive cervical carcinoma than the prototype. Cancer Res *58*, 829-833.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821-829.

Zhang, C., Peng, L., Zhang, Y., Liu, Z., Li, W., Chen, S., and Li, G. (2017). The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. Med Oncol *34*, 101.

Zhang, G.L., Riemer, A.B., Keskin, D.B., Chitkushev, L., Reinherz, E.L., and Brusic, V. (2014). HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology. Database (Oxford) *2014*, bau031.

Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R., *et al.* (2006). microRNAs exhibit high frequency genomic alterations in human cancer. Proc Natl Acad Sci U S A *103*, 9136-9141.

Zhang, L., Wu, J., Ling, M.T., Zhao, L., and Zhao, K.N. (2015). The role of the PI3K/Akt/mTOR signalling pathway in human cancers induced by infection with human papillomaviruses. Mol Cancer *14*, 87.

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

Zhang, X., Bai, J., Yuan, C., Long, L., Zheng, Z., Wang, Q., Chen, F., and Zhou, Y. (2020). Bioinformatics analysis and identification of potential genes related to pathogenesis of cervical intraepithelial neoplasia. J Cancer *11*, 2150-2157.

Zhang, Y., Koneva, L.A., Virani, S., Arthur, A.E., Virani, A., Hall, P.B., Warden, C.D., Carey, T.E., Chepeha, D.B., Prince, M.E., *et al.* (2016). Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures. Clin Cancer Res *22*, 4735-4745.

Zhang, Z., Yu, J., Li, D., Liu, F., Zhou, X., Wang, T., Ling, Y., and Su, Z. (2010). PMRD: plant microRNA database. Nucleic Acids Res *38*, D806-813.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V.L., Travassos da Rosa, A.P., Guzman, H., Cao, S., Virgin, H.W., Tesh, R.B., and Wang, D. (2013). Identification of novel viruses using VirusHunter--an automated data analysis pipeline. PLoS One *8*, e78470.

Zhao, G., Wu, G., Lim, E.S., Droit, L., Krishnamurthy, S., Barouch, D.H., Virgin, H.W., and Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology *503*, 21-30.

Zhao, X., Liu, Q., Cai, Q., Li, Y., Xu, C., Li, Z., and Zhang, X. (2012a). Dr.VIS: a database of human disease-related viral integration sites. Nucleic Acids Res *40*, D1041-1046.

Zhao, Y., Tang, H., and Ye, Y. (2012c). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics *28*, 125-126.

Zhen, S., and Li, X. (2017). Oncogenic Human Papillomavirus: Application of CRISPR/Cas9 Therapeutic Strategies for Cervical Cancer. Cell Physiol Biochem *44*, 2455-2466.

Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., Fuentes, S., Ling, K.S., Kreuze, J., and Fei, Z. (2017). VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology *500*, 130-138.

Zheng, Z.M., and Wang, X. (2011). Regulation of cellular miRNA expression by human papillomaviruses. Biochim Biophys Acta *1809*, 668-677.

Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. PLoS One *8*, e60234.

Zhou, X., Chen, X., Hu, L., Han, S., Qiang, F., Wu, Y., Pan, L., Shen, H., Li, Y., and Hu, Z. (2010). Polymorphisms involved in the miR-218-LAMB3 pathway and susceptibility of cervical cancer, a case-control study in Chinese women. Gynecol Oncol *117*, 287-290.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011). PHAST: a fast phage search tool. Nucleic Acids Res *39*, W347-352.

Zou, S., Caler, L., Colombini-Hatch, S., Glynn, S., and Srinivas, P. (2016). Research on the human virome: where are we and what is next. Microbiome *4*, 32.

Zuna, R.E., Moore, W.E., Shanesmith, R.P., Dunn, S.T., Wang, S.S., Schiffman, M., Blakey, G.L., and Teel, T. (2009). Association of HPV16 E6 variants with diagnostic severity in cervical cytology samples of 354 women in a US population. Int J Cancer *125*, 2609-2613.

zur Hausen, H. (1991). Viruses in human cancers. Science *254*, 1167-1173.

zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer *2*, 342-350.

zur Hausen, H. (2009). Papillomaviruses in the causation of human cancers - a brief historical account. Virology *384*, 260-265.