Genomic data visualization and interpretation

Malachi Griffith, Obi Griffith, Zachary Skidmore
Evomics, Workshop on Genomics
Jan 11, 2019
Český Krumlov, Czech Republic
Introductions to course instructors

Malachi Griffith, PhD
Assistant Professor of Medicine
Assistant Professor of Genetics
Assistant Director, MGI

Obi Griffith, PhD
Assistant Professor of Medicine
Assistant Professor of Genetics
Assistant Director, MGI

Zachary Skidmore, MSc
Staff Scientist, MGI
GenVisR creator

McDonnell Genome Institute, Washington University School of Medicine
The McDonnell Genome Institute has pursued the field of genomics since inception:
>1000 whole genomes, >5000 exomes, >1000 transcriptomes for dozens of tumor types

- Human Genome Project
- First to sequence and analyze a tumor whole genome sequence (Ley et al, 2008)
- Major contributor to TCGA, PCGP, etc.
- Most comprehensively sequenced single patient tumor ever published (Griffith et al, 2015)
- Early proof-of-principle for cancer precision medicine (Griffith et al, 2016)
- Analysis and tools for first personalized cancer vaccine design in humans (Carreno et al, 2015)
- Many other widely used tools

MGI - 1000+ tumor whole genomes, many more exomes, X10 & NovaSeq will accelerate this!
Overview of lab’s research (griffithlab.org)

• Cancer genome analysis
  • Breast cancer, Liver cancer, Lung cancer, Head and neck cancer, etc.
  • Variant Interpretation
  • Immunogenomics

• Precision medicine for cancer
  • Genomics Tumor Board
  • Case Reports
  • Clinical Trials
  • Personalized Cancer Vaccines

• Education projects
  • RNA-seq analysis and cloud computing (CBW, Toronto)
  • Advanced Sequencing Technologies and Applications (CSHL, New York)
  • Genomic Data Visualization/Interpretation (Physalia Courses, Berlin)
  • Precision Medicine Bioinformatics (PR Informatics, Glasgow)
  • High-Throughput Biology: From Sequence to Networks (CSHL / CBW, New York)
  • Workshop on Genomics (Evomics, Český Krumlov)

• Tool development
Where tools/resources do not exist we build them

- **DGIdb**
  - [www.dgidb.org](http://www.dgidb.org)
  - Search genes for known and potentially druggable interactions

- **DoCM**
  - [www.docm.info](http://www.docm.info)
  - Filter against highly curated set of mutations known to cause cancer

- **CIViC**
  - [www.civicdb.org](http://www.civicdb.org)
  - Identify highly curated summaries of clinical interpretations for variants in cancer

- **pVAC-seq**
  - [https://github.com/griffithlab/pVAC-Seq](https://github.com/griffithlab/pVAC-Seq)
  - Personalize vaccine design

- **GenVisR**
  - [https://github.com/griffithlab/GenVisR](https://github.com/griffithlab/GenVisR)
  - Create genomic visualizations

- **regtools**
  - [https://github.com/griffithlab/regtools](https://github.com/griffithlab/regtools)
  - Identify regulatory variants
Whole genome, exome, transcriptome and other ‘omic’ sequencing allows us to detect and confirm many different variant types.
SNVs, Indels, CNVs, SVs, fusions, LOH, expression changes, methylation changes, and more
Comprehensive and integrative analysis methods are needed.
Personalized medicine requires personalized strategies

Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B-lymphoblastic leukemia

Griffith et al. 2016
Strategies to bring genomics information to bear for as many cancer patients as possible

1. Precision medicine targeting of driver mutations

2. Leveraging passenger variants
   a. Tracking minimal residual disease
   b. Identifying neoepitopes
      • Predicting response to immunotherapy
      • Developing personalized vaccines
Precision medicine targeting of driver mutations

1980s: Development of Targeted Therapies

2000: Human Genome Sequencing Project

2001: FDA Approval for Imatinib in BCR-ABL1 CML

2014: FDA Approval for BRCA Testing

2017: FDA Approval of Pembro for MSI tumors

BRAF -> V600E -> Melanoma -> Predictive -> Vemurafenib
ERBB2 -> Amplification -> Breast -> Predictive -> Traztuzumab
EGFR -> L858R -> Lung -> Predictive -> Erlotinib
ALK -> Fusions -> Lung -> Predictive -> Crizotinib
EWSR1-FLI1 -> Fusions -> Ewing Sarcoma -> Diagnostic
DNAJB1-PRKACA -> Fusions -> fHCC -> Diagnostic
VHL -> Loss of function mutations -> Kidney -> Predisposing
... an increasingly long tail of rare but clinically relevant variants
High-throughput sequencing has been largely automated allowing rapid identification of somatic and germline variants.
Interpretation and visualization of genomic alterations remains the bottleneck for realizing precision medicine.

We created CIViC to address this need - an open knowledgebase and curation system for clinical interpretation of variants in cancer.

www.civicdb.org

Joshua McMichael
Strategies to bring genomics information to bear for as many cancer patients as possible

1. Precision medicine targeting of driver mutations

2. Leveraging passenger variants
   a. Tracking minimal residual disease
   b. Identifying neoepitopes
      • Predicting response to immunotherapy
      • Developing personalized vaccines
Circulating tumor DNA (ctDNA) could allow generalized tracking in any cancer type

Wikimedia Commons (CC BY-SA 4.0)
ctDNA tracking in triple negative breast cancer

Stage 2/3TNBC

Blood for ctDNA
Research biopsy

Blood for ctDNA
Research biopsy

Blood for ctDNA
Research biopsy at surgery

Blood for ctDNA every 6 months X 5 years

Tumor/ctDNA sequencing to identify mutations

Tumor/ctDNA sequencing to identify mutations

Tumor/ctDNA sequencing to identify mutations

Tumor/ctDNA sequencing to identify mutations

Clinical Trial: NCT02124902 (Contact PI: Foluso Ademuyiwa)
1. Precision medicine targeting of driver mutations

2. Leveraging passenger variants
   a. Tracking minimal residual disease
   b. Identifying neoepitopes
      • Predicting response to immunotherapy
      • Developing personalized vaccines
Invoking an adaptive immune response against the tumor (focus on CD8+ T cells)

1. **Tumor** produces a unique peptide corresponding to a somatic **mutation**

2. **Processing** and presentation of the tumor specific peptide

3. **Neoantigen** peptide presented by **MHC**

4. **T cell receptor** that uniquely matches the tumor peptide
Neoepitope characterization workflow
T cell mediated cell death

Singh and Gulley, 2014
Personalized cancer vaccine trials

- Kidney Cancer (n = 15 patients)
  - PolyImmune {Durvalumab (MEDI4736) and Tremelimumab} & Vaccine Orchestrated Treatment for Patients With Advanced/Metastatic Renal Cell Carcinoma (NCT03598816). Collaboration with MedImmune.

- Lung Cancer (n = 15 patients)

- Glioblastoma (n = 30 patients)
  - Neoantigen-based Personalized Vaccine Combined With Immune Checkpoint Blockade Therapy in Patients With Newly Diagnosed, Unmethylated Glioblastoma (NCT03422094)

- Breast Cancer (n = 54 patients)
  - Safety and Immunogenicity of a Personalized Polyepitope DNA Vaccine Strategy in Breast Cancer Patients With Persistent Triple-Negative Disease Following Neoadjuvant Chemotherapy (NCT02348320)

- Follicular Lymphoma (n = 20 patients)

- Prostate Cancer (n = 20 patients)
  - Neoantigen DNA Vaccine in Combination With Nivolumab/Ipilimumab and PROSTVAC in Metastatic Hormone-Sensitive Prostate Cancer (NCT03532217). Collaboration with Bristol-Myers Squibb.

- Pancreatic Cancer (n = 15 patients)
  - Neoantigen DNA Vaccine in Pancreatic Cancer Patients Following Surgical Resection and Adjuvant Chemotherapy (NCT03122106).

- Melanoma (n = 12 patients)
  - Dendritic Cell Vaccination in Patients With Advanced Melanoma (NCT03092453). Collaboration with UPenn/Parker ICI.
Genomic data visualization and interpretation
Genomics research has exploded with the rapid advances in DNA sequencing technologies.
Whole genome, exome and transcriptome sequencing allows us to detect and confirm many different variant types.
Why do we create visualizations of genomic data?

• Data exploration and interpretation of results
  • QC analysis
  • Understanding whether/how an experiment worked
  • Discovery

• Communication
  • Slides for presentations
    • e.g. Keynote, Powerpoint, etc.
  • Figures for publications
    • e.g. PDFs, PNGs, etc.
    • Illustrator, Gimp, Inkscape, etc.
    • R and R Studio
Genome browsers
Genome browsers - Ensembl

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes
- Human (GRCh38.p10)
- Mouse (GRCm38.p5)
- Zebrafish (GRCz10)

All genomes
- Select a species

View full list of all Ensembl species
Genome browsers - UCSC

- Genome Browser
  - interactively visualize genomic data
- BLAT
  - rapidly align sequences to the genome
- Table Browser
  - download data from the Genome Browser database
- Variant Annotation Integrator
  - get functional effect predictions for variant calls
- Data Integrator
  - combine data sources from the Genome Browser database
- Gene Sorter
  - find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBiB)
  - run the Genome Browser on your laptop or server
- In-Silico PCR
  - rapidly align PCR primer pairs to the genome
- LiftOver
  - convert genome coordinates between assemblies
- VisiGene
  - interactively view in situ images of mouse and frog

More tools...
Genome browsers - IGV
Fundamentals of data visualization
Fundamentals of data visualization

• Where to learn more about the art and science of visualization:
  • Collection of 40 Nature Methods articles on data visualization: “Points of View” Articles
  • Visual design principles lecture
Which series is more effective (top or bottom)?
Which series is more effective (top or bottom)?

one color dominates
difficult to distinguish
murky
recolored with Brewer palettes

Martin Krzywinski
Which is more effective (left or right?)

Martin Krzywinski
Which is more effective (left or right?)

Excellent organization and consistency. Vertical lines cue continuity. Good use of color.

Chartjunk plentiful. Screaming ornamental and redundant elements. Text inconsistent and illegible.


Which is more effective (left or right?)

[Graph showing cancer types and their frequencies]

_Martin Krzywinski_
When to use a pie chart

Pie charts are good at precisely showing 1:3 proportions.

...but not if the slice is rotated.
When to **not** use a pie chart

- Hard to judge proportions
- Poor use of color
- Hard to read labels
- Over $\frac{1}{2}$ of the categories had to be broken out of the pie chart


Martín Krzywinski
Same data with a redesigned approach
Selected articles on fundamentals of data viz

- Visualizing samples with box plots
- Circos plots
- When to use (and not use) pie charts

- Resources for choosing colors
  - http://colorbrewer2.org/
  - http://mkweb.bcgsc.ca/color/
  - Understanding and using Color Palettes
  - Color palettes for color blindness
  - Names for >9000 colors
    - Including 40 beer colors

- Credit to Martin Krzywinski for his extensive work in this area and many of the above resources
Best practices in visualization
Best practices from this workshop

- *Always* label axes
- Consider readability of font size
- Avoid vertical or angled text if possible
- Avoid unnecessary use of color, point shapes, etc.
- Chose colors wisely
- If individual data points are being plotted and have started to really pile up on top of each other consider using a density function
- Always be transparent about what data manipulation is taking place (e.g. log scale, filtering of outliers, etc.)
Ten Simple Rules for Better Figures (Rougier et al. 2014):
Scientific visualizations should act as a “a graphical interface between people and data”. Try to follow these rules.
1. Know Your Audience
2. Identify Your Message
3. Adapt the Figure to the Support Medium
4. Captions Are Not Optional
5. Do Not Trust the Defaults
6. Use Color Effectively
7. Do Not Mislead the Reader
8. Avoid “Chartjunk”
9. Message Trumps Beauty
10. Get the Right Tool
Example visualizations using R
Histogram and scatterplot to define cell populations

Proportion

Read coverage (X)

Tumor variant allele frequency

WGS data

NF1

Tumor variant allele frequency

Malachi Griffith <mgriffit@genome.wustl.edu>
Clustering to define distinct cell populations
Using a “cell map” to represent the variant clusters in 100 hypothetical cells of a tissue
A ‘fish’ plot is used to represent sub-clones lost and gained over time.
Visualizing copy number variation

Mean tumor-normal depth per 10kb window plotted

KRAS amplification in a metastatic breast cancer
A ‘waterfall’ plot is one way to visualize the pattern of variant recurrence in a cohort of samples.

https://github.com/griffithlab/GenVisR
Heatmaps are a common way to simultaneously visualize multiple features of a dataset.
GenVisR was created to help others make common genomic visualizations.
Acknowledgements: Griffith Lab group members

Malachi Griffith  Obi Griffith  Benjamin Ainscough  Erica Barnell  Katie Campbell  Kaitlin Clark  Adam Coffman  Kelsy Cotto

CIViC/Regulome  Regtools/pVA C-seq  Lymphoma  pVAC-seq  pVAC-seq/CIViC  CIViC/Lymphoma  CIViC  HCC

Arpad Danos  Yang-Yang Feng  Felicia Gomez  Jasreets Hundal  Susanna Kiwala  Kilanin Krysiak  Lynzey Kujan  Jason Kunisaki

CIViC  CIViC Lymphoma  genVisR HCC/SCLC  CIViC/Komen  HCC/SCLC  DGIdb SCLC  Director Informatics

Josh McMichael  Cody Ramirez  Zachary Skidmore  Nick Spies  Lee Trani  Alex Wagner  Jason Walker
Acknowledgements: McDonnell Genome Institute

McDonnell Genome Institute @ Washington University School of Medicine

Introduction to GenViz course site

www.genviz.org