ADVANCED VISUAL ANALYSIS OF GENOMIC VARIATIONS WITH NGB -- NEW GENOME BROWSER

Gennadii Zakharov

August, 2017
Variations

- Variations -- differences between reference genome and particular sample
- Easier to store and analyze
- Variations types:
  - SNP
  - Short INDELs
  - Structural variations
- The larger variations usually are harder to detect and analyze
Variation analysis pipeline

- Raw reads trimming/processing
- Align
- Alignment postprocessing (Realignment, recalibration)
- Variant calling
- Variant annotation

- Manual data review:
  - Required in clinical diagnosis
  - Required for non-obvious cases in researches
### Why a new genome browser?

**Desktop applications**
- Wide set of functions / features
- Good performance
  - Needs installation and version updates
  - Hard to invoke via a Web-link
  - Low Web/cloud support
- Leader - **IGV**

**Web-based applications**
- No need to install / update
- High Web/cloud support
  - Much less functionality
  - Poor/low performance
    - A lot of limitations to keep acceptable speed
- Leader - **Jbrowse**

#### New Genome Browser
- Wide set of functions / features
- High performance
- No need to install / update
- High Web/cloud support

<table>
<thead>
<tr>
<th>Functionality + Speed</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web/Cloud enablement</td>
<td>High</td>
<td>JBrowse</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>IGV</td>
</tr>
</tbody>
</table>
New genome browser (NGB)

- Web at a speed and functionality of desktop
- Provides unique features, compared to other solutions:
  - Big / cloud data excellent performance
    - Fast access to huge (500 Gb and more) data files
    - Easy access to cloud data (Amazon S3, Hadoop, etc.)
  - Most advanced analysis of mutations
    - Unique SV (Structural Variations) visualization
    - Public and in-house databases integration (mutation effects annotations)
- Fast and flexible feature extension by EPAM team

https://github.com/epam/NGB
NGB provides three installation options:
- Jar file
- War file
- Docker Image

<table>
<thead>
<tr>
<th>Data source</th>
<th>Supported file types/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server local filesystem</td>
<td>All file types Reference data (FASTA + GFF/GTF) can only be accessed like this</td>
</tr>
<tr>
<td>NFS, FTP, HDFS storages mounted to Server local filesystem</td>
<td>All file types</td>
</tr>
<tr>
<td>AWS S3</td>
<td>Remote access to BAM files Currently access tokens to S3 bucket should be stored on NGB server (database). Also signed URLs can be used</td>
</tr>
<tr>
<td>Generic HTTP URL</td>
<td>Remote access to BAM, VCF, BED files Files of other types would be downloaded to the server during registration</td>
</tr>
<tr>
<td>FTP (direct access)</td>
<td>Remote access to BAM, VCF, BED files Files of other types would be downloaded to the server during registration</td>
</tr>
</tbody>
</table>
NGB FEATURES
Browsing files

Open files via URL or from server file system

Explore registered datasets and files
Tracks: VCF, BAM/CRAM, WIG

Reference (Fasta files)

Supported tracks:
- Gene (GFF/GTF files)
- Alignment (BAM/CRAM files)
- Variations (VCF files)
- BED
- Wig/BedGraph
- SEG

Each file is represented as a single data track.
Tracks: Genes (GFF/GTF), BED, SEG
Genes visualization
**Variations table**

- Display list of variations loaded from the project's VCF files
- Extract gene information form ANN field
- Group variations by type
- Sort and filter by any field from initial VCF file - population frequency, effect predictor score, ...
- Navigate to variation by single click

<table>
<thead>
<tr>
<th>Type</th>
<th>Chr</th>
<th>Gene</th>
<th>Position</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUP</td>
<td>16</td>
<td>TSC2</td>
<td>2080935</td>
<td>i</td>
</tr>
<tr>
<td>DEL</td>
<td>20</td>
<td>MACROD2</td>
<td>14766433</td>
<td>i</td>
</tr>
<tr>
<td>DEL</td>
<td>7</td>
<td>ISPD, ISPD-AS1</td>
<td>16257894</td>
<td>i</td>
</tr>
<tr>
<td>DEL</td>
<td>21</td>
<td>CHODL, AP000998.2</td>
<td>18065750</td>
<td>i</td>
</tr>
</tbody>
</table>

![Variations table image]
Variations filtering

<table>
<thead>
<tr>
<th>DATASETS</th>
<th>VARIANTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Chr</td>
</tr>
<tr>
<td>BND</td>
<td>13</td>
</tr>
<tr>
<td>BND</td>
<td>13</td>
</tr>
<tr>
<td>BND</td>
<td>13</td>
</tr>
<tr>
<td>INV</td>
<td>13</td>
</tr>
<tr>
<td>BND</td>
<td>13</td>
</tr>
<tr>
<td>BND</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANNOTATION</th>
<th>Annotation</th>
<th>Annotation_D</th>
<th>Gene_Name</th>
<th>Gene_ID</th>
<th>Feature_Type</th>
<th>Feature_ID</th>
<th>Transcript_Type</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRON_VARIANT</td>
<td>MODIFIER</td>
<td>R1</td>
<td>ENSG00000139687</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRON_VARIANT</td>
<td>MODIFIER</td>
<td>R1</td>
<td>ENSG00000139687</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Variations visualization

- Variations type indication
- Variation info in popup
- Variation detailed description
- All data from VCF are parsed and displayed in variation info
Alignment – BAM & CRAM

- Fast alignment visualization
- Read tooltip and detailed info
- Collapsed view for high-coverage samples
- All sorting grouping and coloring reads modes from IGV: by read strand, pair orientation, insert size, first in pair strand, etc.
- Softclipped bases
Databases Integration

- Genes/proteins information from ENSEMBL, UniPROT
- Variations data form dbSNP
- Other internal/external databases (by request)
RNASEq

- Splice junctions
- Shorten introns (exome only) view hides all introns and shows only a part of a reference that corresponds to gene exons
Additional Features

- Search and navigation
- Bookmarks
- Collaborative work
- Session sharing
- Shortening introns mode
- Interface customization
- Embedding
- Structural variations support
CASE STUDY ANALYSIS USING NGB
SNP

- View supporting reads in alignment
- Short Variation info
- Amino acid change in visualizer
- Missence SNP
INDELs analysis

- Supporting reads
- Short Variation info
- Mismatched letters - soft clipped parts of read
- View amino acid change in visualizer
- In frame deletion
• Genomic rearrangements that effect more than 1 Kb (https://www.ncbi.nlm.nih.gov/dbvar/content/overview/)
• Typically affects a sequence length about 1Kb to 3Mb
• Approximately 13% of the human genome are defined as structurally variant in the normal population
• Balanced: inversions and translocations
• Imbalanced: Duplications, insertions and deletions
• Identified using coverage, split reads, mate reads, ...

• Full SV support according to VCF 4.3 specification
• Displaying variation type and additional data
• Variations in split screen
• Variations effect visualizer
  – Provides almost publication ready visualization for variations.
  – Provide information on a resulting protein product, including exons/domains, that were affected by a variation
Inversion

- Alignment and variation on the both sized on inversion
- Reads supporting reads with abnormal insert size on both ends
- Inversion cause two fusion proteins
- Reading frame and direction remains the same
- The resulting fused protein has ALK kinase domain synthesized from different promotor
- Potential oncogenic SV
NGB provides unique and publication ready visualization for SVs. It provides information on a resulting protein product, including exons/domains affected by a variation.

**Translocation (BND)**

**Inversion**

**Duplication**
Chromosomes fusion

• See other side of breakpoint in variation mark
• Supporting reads - grouped by chromosome of mate
• Reading frame and direction remains the same
• Oncogenic SV
Duplication

- View both variation ends in split screen
- Supporting reads with abnormal orientation
- Incorrect duplication start
The NGB visualization of the FGFR3-TACC3 fusion is shown in Figure 5. Unlike the Svviz plot, the visualization is fully interactive HTML5 in the browser.

Red highlighting is used to show the breakpoints relative to the coding regions in the alternative allele view and the red line shows the fusion points in the reference allele view.

Figure 5. FGFR3-TACC3 tandem duplication fusion exon level visualisation in the New Genome Browser. Protein domains and exons affected by the structural variant are highlighted in colours.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382922/
NEW GENOME BROWSER + MOLECULAR VIEWER

Embedded molecular viewer enables showing proteins related to the gene shown in genome browser; highlights regions of interest - mutations, domains, etc.
NGB Use Case #1 – Pharma company

Two instances + one coming soon. 50+ users and up to 100 by Q4-2017

“Thank you for your effort, the team was really impressed with the quality of the product developed in such a short period of time!”
... Software Development Manager

“I want to congratulate all the team involved in the development of NGB. I think this genome browser is really cool!”
... Bioinformatics Senior Scientist

“I’m totally amazed by the NGB - absolutely fantastic work. This will definitely be super useful for us.”
... Principal Informatics Scientist

“Unlike other genome browsers, NGB is fast, works in Chrome browser and can be used on any computer without need for ancillary software or setup. For me, the data visualizations are superb and very clear - with formatting familiar to those used to using IGV. Coupled with the “Variants” table it enables rapid review of filtered subsets of variants from internal and external datasets. I also love how I can share views of specific regions of datasets with colleagues (difficult to do this any other way other than static screengrabs).”
... Translational Genomics Lead
NGB Use Case #2 – Animal health company

One instance to be launched in Q3. Up to 100 users by Q4-2017

“Awesome!”
... - Senior Scientist Bioinformatics
NGB use case #3 – Hospital
One instance to be launched in Q3. Up to 50 users by Q4-2017

RNA-Seq Analysis:
- Deep coverage
- Useful features - grouping, sorting, coloring
- Splice junctions
- Shorten introns

“It's looking very good...We are ready to try and incorporate this into our clinical workflow as your team develops more.

I love the ‘shorten introns’ feature

I am impressed that you already have a visualizer for structural variants

I am impressed with your team’s work so far and look forward to working together.”

... Assistant Professor of Pathology
Thank you for your attention!