

BioPivot: Applying Microsoft Live Labs' Pivot to Problems in Bioinformatics

Stephen Taylor, CBRG

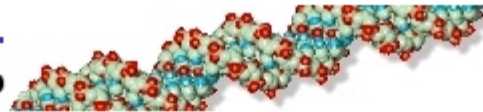
GMOD Europe 2010

Introduction

- Visualization of large numbers of genome regions
- Querying and filtering properties of genome regions
- Pivot and BioPivot tools
- Open discussion of other applications of technology

CBRG

Computational **B**iology
Research **G**roup



Introduction

You are here: CBRG home

CBRG Home

- CBRG personnel
- computing support
- contact details
- search CBRG web site

CBRG Accounts (molbiol)

Analysis tools

Training courses

Tutorials

Unix help

Examples

Papers

Collaborative data

Presentations

Oxford-only section

FAQ: CBRG + UNIX

FAQ: Bioinformatics

Links

Welcome to the CBRG

The Computational Biology Research Group (CBRG) provides [computing support](#) for bioinformatics analysis at the University of Oxford. We have expertise in many aspects of bioinformatics (sequence analysis, microarrays, proteomics and integration). We especially encourage collaborations that require writing custom software, bioinformatics tools and databases.



An account with the CBRG has many [benefits](#) and gives automatic access to a large number of molecular biology computing packages and to numerous biological databases.

We are based at the [Sir William Dunn School of Pathology](#) and at the [Weatherall Institute of Molecular Medicine](#). Full details can be found on the [contact details](#) page.

Bioinformatics analysis tools online:

A wide range of bioinformatics programs are available online via [EMBOSS Explorer](#). You will need a molbiol username and password to use these tools.

Other web tools available include: [BLAST](#), [MASCOT](#) and [BASE](#) (see [full list](#)).

Over 50 different GBrowse databases

- Many labs started wanting GBrowse
- Human
- Mouse
- Bacterial

- Time series
- Arrays
- ChIP-Seq
- RNA-Seq

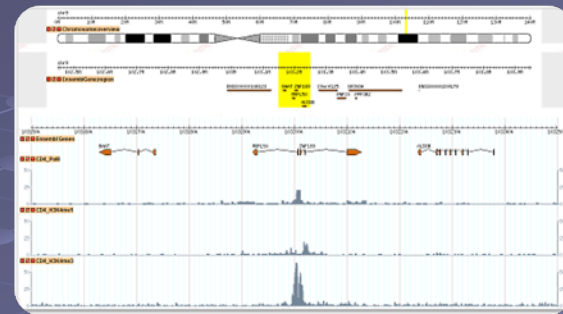
Next Generation Sequencing

- Histone modification Data
- ChIP-Seq
- Interaction cis/trans data
- PCR amplified regions
- RNA-Seq
- Exome Sequencing/SNP detection

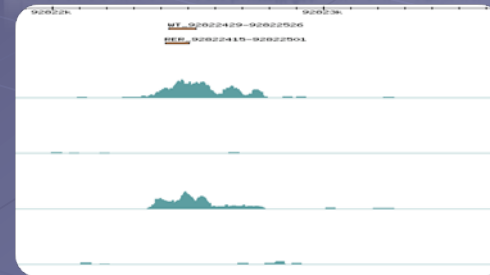
ChIP-Seq example

NGS reads

Map



Peak pick



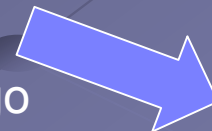
Extract sequences from features



Motif extract



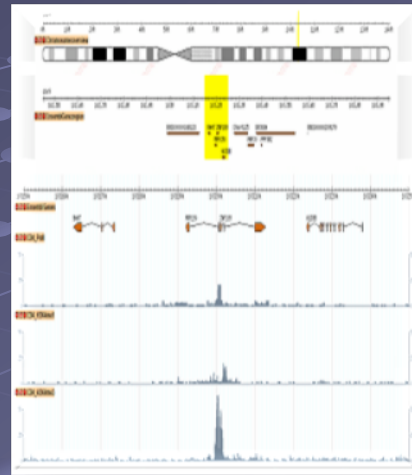
Weblogo



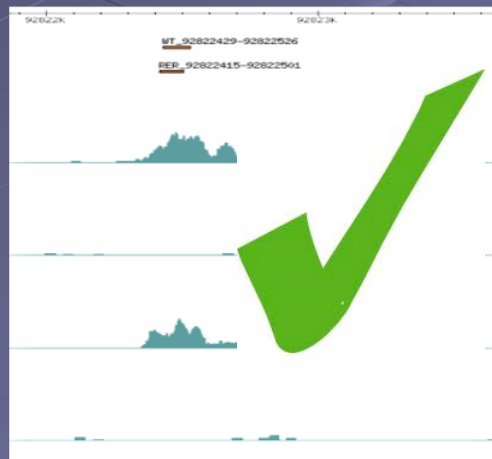
ChIP-Seq

NGS reads

Map



Peakfind



Problems

- Experimental conditions
- Antibody
- Peak finders give false positives
 - Lots of parameters
 - Must choose a suitable cut-off
 - Eyeballing lots of peaks

Further Analysis

- Which of my peaks overlap with:
- Genes
- Exons
- Promoters
- CpG Islands
- Areas of conservation
- etc

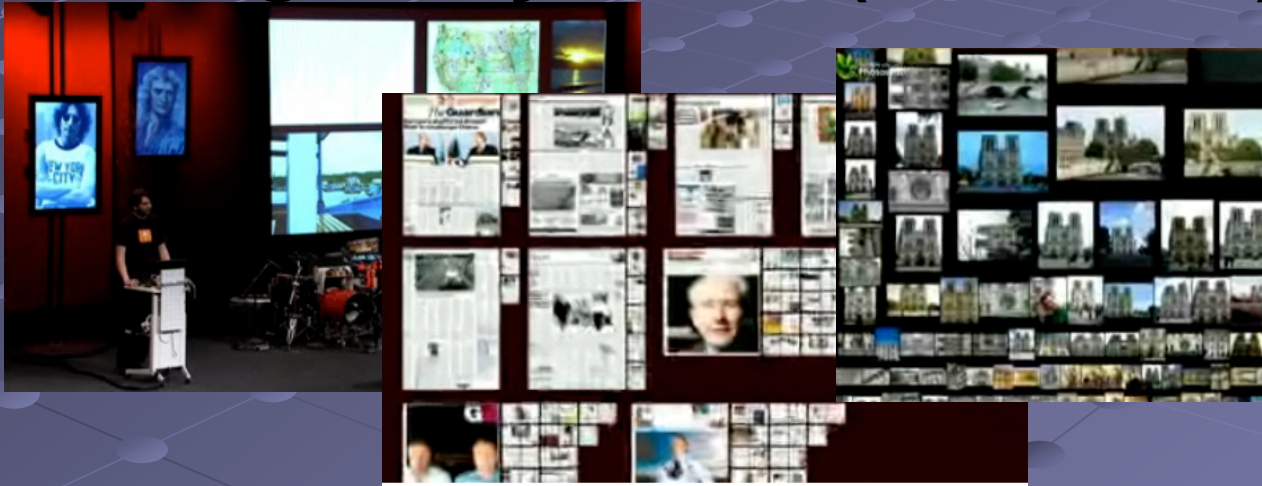
Traditionally

- Make spreadsheet of data with links to Gbrowse/UCSC regions of interest
- Click/Filter various parameters
- Add data to spreadsheet each new analysis

SLOW!
BORING!

Deep Zoom Tech

● Blaise Aguera y Arcas (TED 2007)



● Seadragon/Photosynth

- <http://www.seadragon.com/showcase/>

● Microsoft Live Labs' Pivot

- <http://www.getpivot.com/>

Wouldn't it be cool...

- To use this in bioinformatics...?
- Take thousands of regions of interest of genome
- View and Filter seamlessly on metadata

BioPivot Tools

- GFF3 of ROI
- Ninth column contains 'facets'
- Choose your GBrowse or UCSC Browser View
- Run the command:

```
gff32pivot my.gff3 -dzi -generateimages -conf mytypes.cfg  
-o my.cxml -browser gbrowse2
```

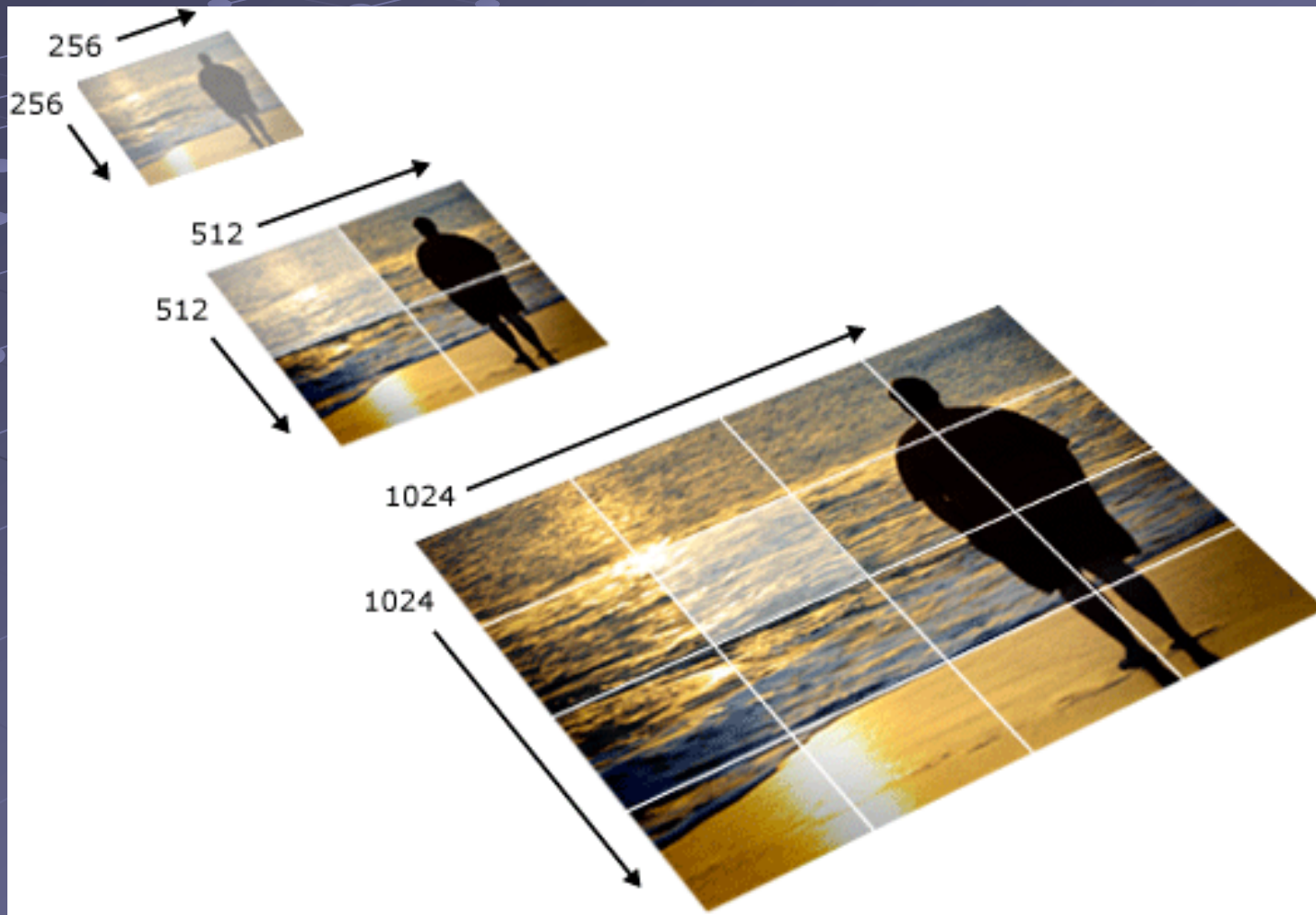
BioPivot Tools

- Parsers for peakfinders
- Annotate a GFF3 file
 - nearest gene
 - exons, introns, intergenic, intragenic
 - TSS/TES up and down stream regions
- Overlaps of GFF3/BED features

Open Source

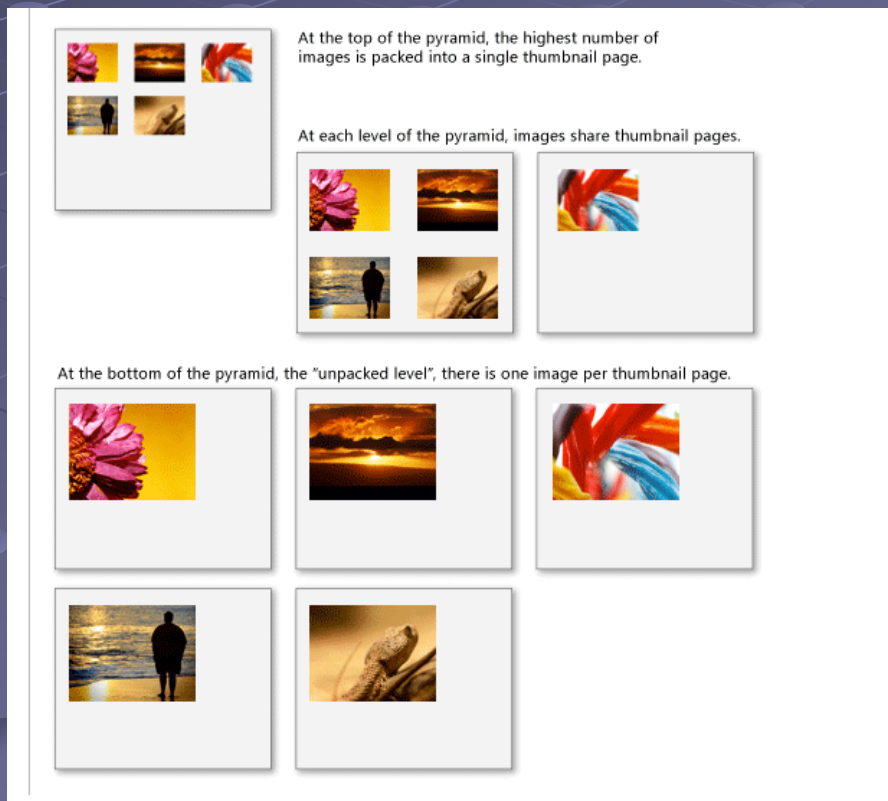
- Zoomable User Interfaces (ZUIs)
- OpenZoom
 - <http://openzoom.org/>
- SDK for Flash, Flex & AIR
- APIs

Deep Zoom Image



Deep Zoom Collection

Groups of tiled images



CXML File

```
<-Collection SchemaVersion="1.0" Name="example.annotated.overlap.gff3">
-<FacetCategories>
  <FacetCategory Name="Alias" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="GO_BP" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="GO_CC" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="GO_MF" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="ID" Type="Number" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="alias_link" Type="Link" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="chromosome" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="distance_from_exon1" Type="Number" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="end" Type="Number" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="location_overlap" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="overlap_CPG" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="phase" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="refseq" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="refseq_strand" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="score" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="source" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="start" Type="Number" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="strand" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
  <FacetCategory Name="type" Type="String" p:IsFilterVisible="true" p:IsWordWheelVisible="true" p:IsMetaDataVisible="true"/>
</FacetCategories>
-<Items ImgBase="img/images.xml">
-<Item Id="0" Img="#0" Name="1">
  -<Facets>
    -<Facet Name="Alias">
      <String Value="ZNF189"/>
    </Facet>
    -<Facet Name="GO_BP">
      <String Value="negative regulation of transcription from RNA polymerase II promoter"/>
      <String Value="regulation of transcription"/>
      <String Value="DNA-dependent"/>
    </Facet>
    -<Facet Name="GO_CC">
      <String Value="intracellular"/>
      <String Value="nucleus"/>
    </Facet>
    -<Facet Name="GO_MF">
      <String Value="transcription factor activity"/>
      <String Value="zinc ion binding"/>
      <String Value="metal ion binding"/>
    </Facet>
    -<Facet Name="ID">
      <Number Value="1"/>
    </Facet>
  </Facets>
</Item>
</Items>
</Collection>
```

To Do

- Installation scripts
- Deploy in a web browser using Silverlight
- RNA-Seq parsers e.g. cufflinks, DESeq
- Get feedback from the community
- What else can we do with this tech?
- <http://www.cbrg.ox.ac.uk/data/biopivot>

Acknowledgements - Code

● OpenZoom

- <http://openzoom.org/>

● Cisgenome

- <http://www.biostat.jhsph.edu/~hji/cisgenome/>

● BEDtools

- <http://code.google.com/p/bedtools/>

Acknowledgements - People

- Jim Hughes
- CBRG Team
- GMOD Team