Gene prioritization Strategies

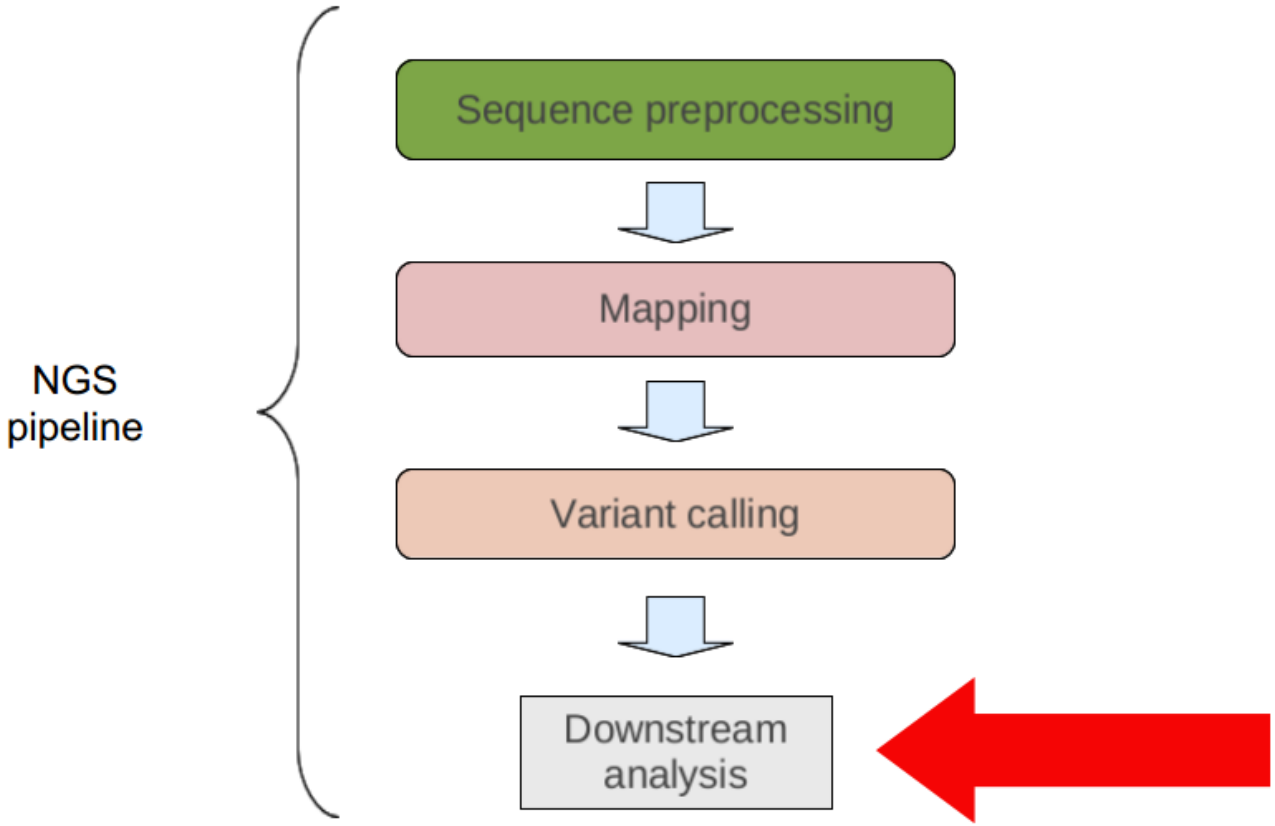
Luz Garcia-Alonso

Valencia, July 2012

1. ROADMAP
2. PROBLEM: Large number of candidate genes/variants
3. PRIORITIZATION
4. EXISTING METHODS
5. FUTURE WORK

ROADMAP

OBJECTIVE: To find the variants associated to the phenotype under study



ROADMAP

OBJECTIVE: To find the variants associated to the phenotype under study

NGS pipeline

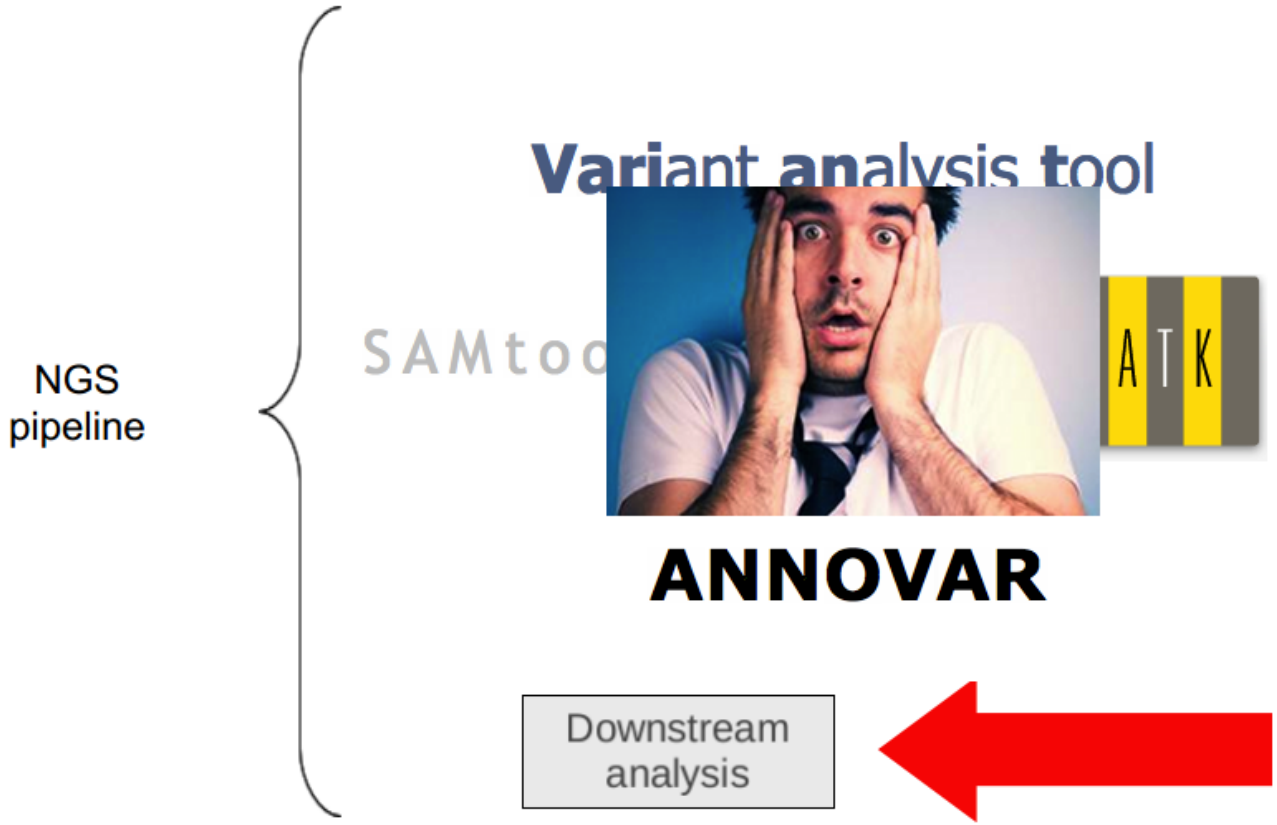
Variant analysis tool

SAMtools

GATK

ANNOVAR

Downstream analysis



The candidate gene lists generated contain hundreds of genes among which only one or a few are of interest

PROBLEM: Large number of candidate genes/variants

An end has a start ...

PROBLEM: Large number of candidate genes/variants

- The experimental validation of every candidate

PROBLEM: Large number of candidate genes/variants

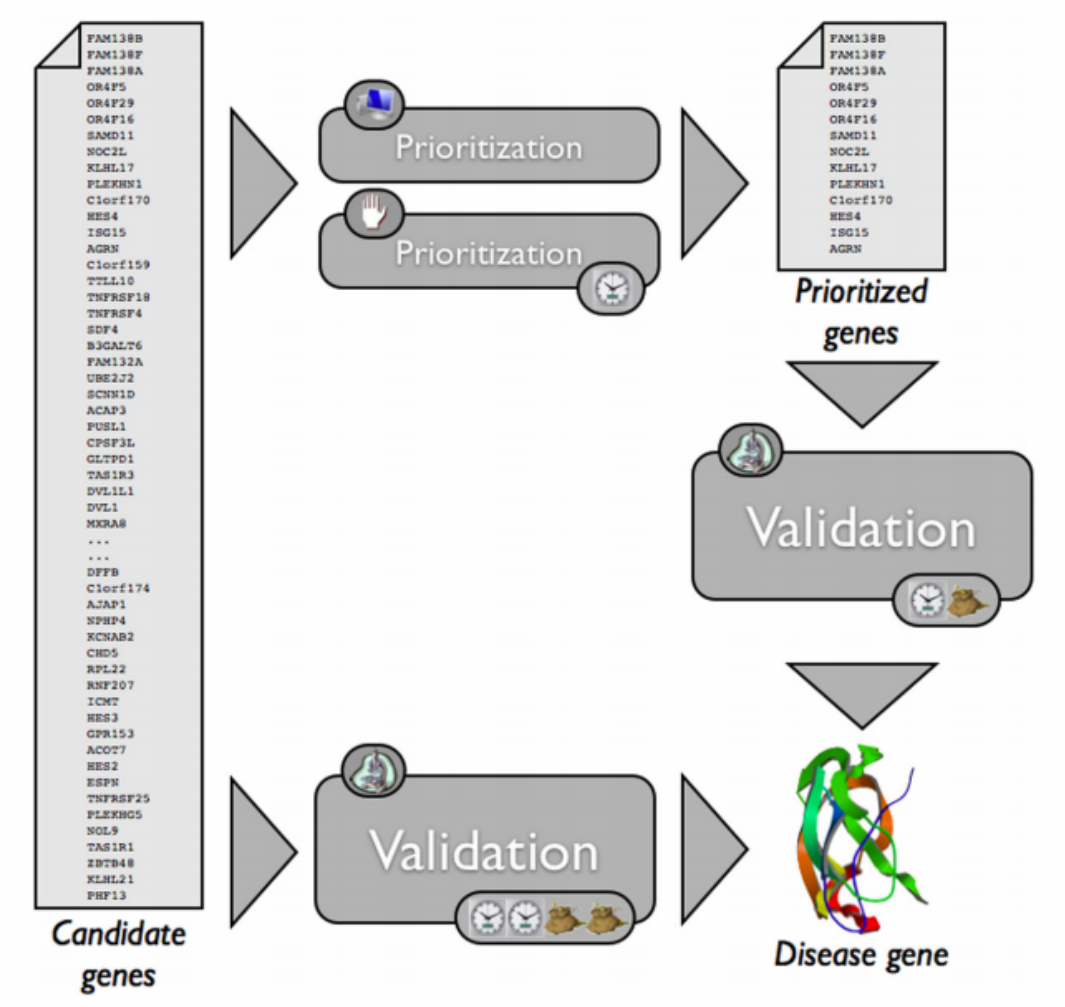
- The experimental validation of every candidate
- **BUT** it is expensive and time consuming.

PROBLEM: Large number of candidate genes/variants

- The experimental validation of every candidate
- **BUT** it is expensive and time consuming.
- It is needed to prioritize the candidate genes using a computational approach at almost no cost and to experimentally validate only these genes.

PRIORITIZATION

The identification of **the most promising** genes among a list of candidate genes.



Number of genes implicated

1. Mendelian or complex disease?
2. Rare alleles or multiple common variants?
3. Family-specific variant?
4. High or small effect (penetrance)?

Biological profile

1. Is any biological process known to be implicated?
2. There are known disease-assocated genes?

IMPORTANT Different methods, different hypothesis tested!

Looking for ONE gene ...

ENDEAVOUR: based on how similar a candidate gene is to a profile derived from genes already known to be involved in the processes

`http://homes.esat.kuleuven.be/~bioiuser/endeavour/tool/endeavourweb.php`

Pros

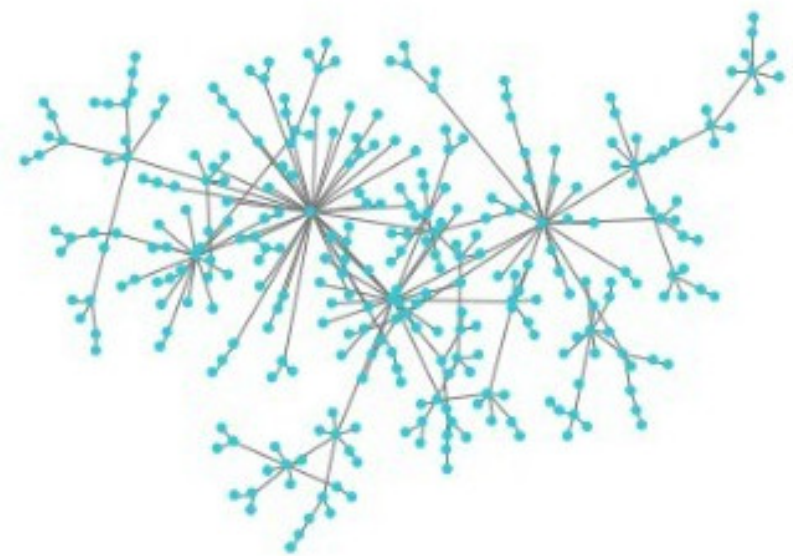
- Eassy to use
- Several biological databases screened

Cons

- Reference genes needed
- Does not take into account the cell complexity
- Does not take into account colective effects of candidate genes
- The genes more annotated are more likely to have a good score

Why protein networks?

- Disease-associated variants occur more frequently in protein-coding regions than expected
- The integration of the whole set of protein interactions provides detailed map (network) about the pathways and molecular complexes and brings a safe place to work with.
- Non-random placement of disease-causing genes in the network
- A network-neighbour of a disease-causing gene is likely to cause a related phenotype



NetworkPrioritizator: based on network location similarity of candidate gene whit respect to reference genes.

Pros

- Takes into account the whole cell complexity
- Several biological databases screened

Cons

- Reference genes needed
- Does not take into account colective effects between candidate genes
- The biological processes understudied are less likely to be well prioritized

Looking for SEVERAL genes ...

NetworkMiner: based on subnetwork agregation between the candidate genes. Finds significant subnetworks of protein-protein interactions within a list of ranked genes/proteins
`http://babelomics.bioinfo.cipf.es/functional.html`

Pros

- Takes into account the whole cell complexity
- Takes into account colective effects between candidate genes
- Does not need reference genes, but they are allowed

Cons

- The biological processes understudied are less likely to be well prioritized
- Valid for complex phenotypes were several genes are expected to be associated

jFamNet: Some human diseases are known to cluster in families. That is, we can expect each family has a different affected gene but located in the network neighborhood.

Pros

- Takes into account the whole cell complexity
- Takes into account colective effects between candidate genes
- Does not need reference genes, but they are allowed

Cons

- Several families needed
- The biological processes understudied are less likely to be well prioritized

FUTURE WORK

1. Refine the previous methods
2. Moving to a high resolution map of the cell able to work with variant level
3. Including genome and transcriptome regulation network
4. Develop new methodologies able to test new hypothesis