gemini: a scalable, agile framework for mining genome variation

Uma Devi Paila and Aaron Quinlan
Department of Public Health Sciences
Center for Public Health Genomics, University of Virginia.

cphg.virginia.edu/quinlan/
github.com/arq5x/gemini

Motivation

★ Tools for population-scale studies of genome-wide variation require **efficient** analytical **frameworks**.
★ Existing approaches are ill-suited to the analysis of variation in the context of **genome annotations**.
★ Researchers need a simple, efficient, and flexible framework for large-scale data exploration.

Enhanced SQL engine allows **selection** and **filtering** based on individual genotypes. **Scales to 1000s of samples.**

```
SELECT chrom, start,
gene, impact, hwe,
in_dbsnp, in_omim,
gts.NA12878
FROM variants
WHERE is_lof = 1
AND aaf <= 0.01
AND call_rate \
    >= 0.95
```

Ad hoc queries and filters using custom SQL engine

```
> gemini query -q [QUERY] my.db
```

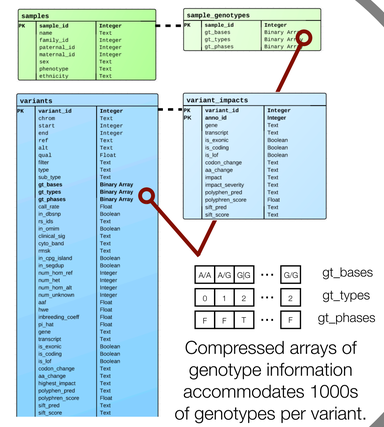
Genome annotations place variants in context

OMIM, COSMIC, dbSNP, CpG, Repeats, SegDups, Conservation, (more), DNase, FAIRE, TF Binding, Recombination, ESP

snpEff
VEP
VAT*

VCF file

PED+ file
(sample relationships)

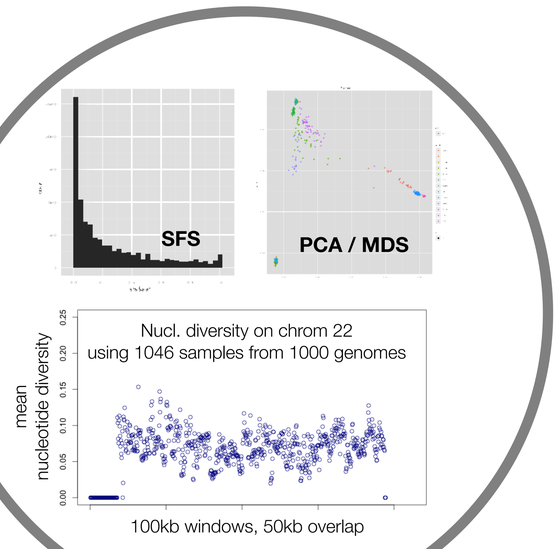


Compressed arrays of genotype information accommodates 1000s of genotypes per variant.

Load VCF into gemini database framework.

Variants automatically documented w/ large set of genomic annotations
Useful PopGen metrics pre-computed
Fast 100,000 genotypes per second

```
$ gemini load -v my.vcf -t VEP my.db
```



"Shortcuts" for common analyses, QC, **visualization**

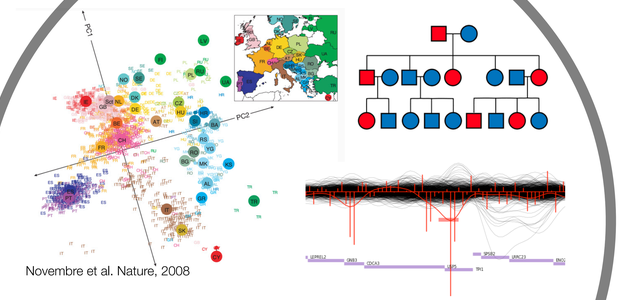
```
> gemini stats --sfs my.db
> gemini stats --mds my.db
> gemini stats --tstv my.db
> gemini stats --vars-by-sample my.db
> gemini stats --gts-by-sample my.db
```

*VAT to be supported in summer 2012

gemini's strengths

★ **Flexible.** Employs an agile database design and annotation system
★ **Scalable.** Accommodates genome-scale variation for 1000s of samples
★ **Powerful.** Employs an agile database design and annotation system
★ **Useful.** Simple interface allows rapid analyses and new tool devel.
★ **Informative.** Place variation in the context of a wealth of genome annot.

A framework for medical and population genetics



- Mendelian disorders
- Disease associations
- Family-based studies
- Cancer genomic studies

```
# gemini imports
import gemini_utils as util

def my_tool(c, args):
    """
    Execute a query against the gemini database and
    conduct a custom analysis on the results
    """
    # build and execute the relevant query against the da
    query = "SELECT * FROM variants \
             WHERE is_coding = 1"
    c.execute(query)

    # loop through the results.
    for row in c:
        gt_types  = np.array(unpack(row['gt_types']))
        gt_phases = np.array(unpack(row['gt_phases']))
        gt_bases  = np.array(unpack(row['gts']))

        # MAGIC HAPPENS HERE

def run(parser, args):
    if os.path.exists(args.db):
        conn = sqlite3.connect(args.db)
        conn.isolation_level = None
        conn.row_factory = sqlite3.Row
        conn.cursor()
```

- Screen for compound hets
- "Windowing" analyses
- Collapsing methods
- Fst | Hz | ...

A simple, efficient framework/API for exploratory analysis and new tool development

```
> gemini comp_hets -lof my.db
> gemini window -pi_hat -w 1000 my.db
> gemini my_tool -arg1 -arg2 my.db
```