Sequencing Populations to Find Causal Genetic Variants
ICRISAT

Richard Mott

University College London, UK

September 20, 2017
Why Sequence Populations?

- Catalogue Genetic Variation
- Genome Wide Association Studies (GWAS)
- Structural Variation
- Population Demography and History
- Signatures of Selection and Drift
Sequence a small(ish) number of individuals from representative populations and catalogue all sequence variants

- Phase variants to create a **haplotype reference panel, HRP**.
- Most common variants (freq > 1%) will segregate in a HRP of just a few hundred samples.
- The number of common haplotypes at any locus is small(ish) (< 100).

Genotype a different (large) sample in a case-control study at ~ 1 million SNPs using arrays.

Use the fact that sample chromosomes are mosaics of the HRP to **impute** all the variants in the panel into the sample.

Perform a GWAS on the imputed genotypes.
Imputation - Predicting Missing Genotypes

From Marchini and Howie, 2010 Nat Rev Genetics.
What about other Species or non-standard Populations?

- Only have a reference genome
- No catalogue of variation available
- No genotyping arrays
- Subpopulation that is very different from reference.
Next-Generation Sequencing

- High coverage sequencing of a few(ish) individuals
  - 100,000 Genomes
  - \( \sim 30x \) read coverage
- Low-coverage sequencing of a population sample.

Three samples from the CONVERGE study of major depression, sequenced at \( \sim 1.5x \)
Low-Coverage Sequencing

- **Disadvantages:** Many loci absent/incompletely covered.
- **Advantages:**
  - Cheap
  - Multiple SNPs on same read give haplotype data.
  - Pileup of all reads identifies most segregating variants - no need for haplotype reference panel
Assume population descends from $K$ unknown founder haplotypes

- $K$ is small, typically 2 – 60.
- $K$ depends on population history (was it bottlenecked?).
- Each chromosome of the population is a mosaic of the founders

Sequence many individuals from population using short reads

- Coverage depends on the numbers of samples and founders.
- Sample Size depends on size of genetic effects to be detected.

Have to estimate both the founder haplotype space and the chromosome mosaics in order to impute each individual’s genotypes.
To a first approximation, genetic effects depend additively on the number $x = 0, 1, 2$ of reference alleles, the dosage.

$$y_i = \beta_g x_{gi} + e_i$$ for individual $i$ at SNP $g$. 

![Graph showing the relationship between dosage and phenotype with p-value 1.350e-03]
Genotype Dosages

- Imputation estimates dosages as the **expected number** of reference alleles carried by an individual.

  \[ x_{gi} = \sum_{h,k} \Pr(\text{founders are } h, k \text{ at SNP } g \text{ in individual } i) \times (\Pr(h \text{ is ref at } g) + \Pr(k \text{ is ref at } g)) \]

- Imputed dosages not exactly 0, 1, 2, due to errors in founder haplotypes, sequencing, read-mapping, structural variants etc.

- Correlation \( R^2 \) with array dosages measures accuracy.
Like many other algorithms, STITCH models the imputation problem as a Hidden Markov Model, but adapted to low-coverage sequencing.

- The Haplotype Mosaics are the hidden states of the HMM
- SNP calls in the sequence reads are the observed data

Input data:
- Reference Genome Sequence
- Low-coverage sequence from a population sample, aligned to the reference

Algorithm:
- Call sequence variants in aligned reads (e.g. by GATK)
- Initialise $K$ ancestral haplotypes randomly.
  - A haplotype emits reference or alt allele at a SNP.
- Iterate $\sim$ 40 times:
  - Find individual haplotype mosaics using current haplotypes.
  - Update haplotype emission probabilities using sample reads and haplotype mosaics.
- After convergence, impute genotypes using haplotype mosaics and haplotype emission probabilities.
STITCH’s Algorithm

Shown are real data from CFW mice with K=4 founder haplotypes across 3kb on chr 19 containing 20 imputed SNPs. Each SNP in the 4 reconstructed haplotypes shown as a vertical bar split proportionally to the probability of emitting the reference (black) or alternate (grey). Sample reads are similarly coloured.
Example 1: CFW Outbred Mice

- CFW are a commercially-available outbred stock of mice
- They are thought to have been bottlenecked to two unknown individuals (4 haplotypes) about 100 generations ago.
- They are maintained in large colonies with effective population size estimated $\sim 1000$.
- They have little population structure.
- Resemble a wild population isolate?
CFW Outbred Mice

- Sequenced 2000 CFW mice at 0.15x coverage
- From a $\sim 350x$ pileup of all reads
  - Identified 5.7 million high-quality SNPs,
  - Of which 98% are present in sequenced inbred strains.
- These SNPs are tagged ar $R^2 > 0.98$ by 360K SNPs.
- Phenotyped each mouse for $> 100$ traits.
- Mapped 255 quantitative trait loci (QTL) representing 156 unique regions for 92 phenotypes.
255 QTLs for 92 phenotypes
Results are shown for STITCH applied to CFW mice. Validation dataset is the Illumina MegaMUGA array (left) and 10X Illumina sequencing (right). pre-QC=black, post-QC=purple, Beagle (default) (red) and findhap (maxlen=10000, minlen=100, steps=3, iters=4) (green) genome-wide for n=2,073 mice featuring 7.07M SNPs before QC and 5.72M after QC. STITCH is run using K=4, diploid method, 40 iterations. Post-QC is SNPs with info>0.4 and HWE p-value > 1 × 10^-6.
Imputation Accuracy in CFW mice as a function of $K$

<table>
<thead>
<tr>
<th>$K$ haplotypes</th>
<th>$R^2$</th>
<th>Time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.922</td>
<td>14.4</td>
</tr>
<tr>
<td>3</td>
<td>0.875</td>
<td>22.6</td>
</tr>
<tr>
<td>4</td>
<td>0.920</td>
<td>35.9</td>
</tr>
<tr>
<td>4, Read Unaware</td>
<td>0.898</td>
<td>37.7</td>
</tr>
<tr>
<td>5</td>
<td>0.919</td>
<td>50.6</td>
</tr>
<tr>
<td>6</td>
<td>0.918</td>
<td>75.1</td>
</tr>
<tr>
<td>7</td>
<td>0.917</td>
<td>97.9</td>
</tr>
<tr>
<td>8</td>
<td>0.911</td>
<td>127</td>
</tr>
</tbody>
</table>

"Read Unaware" means that haplotype information from SNPs on the same read was ignored by artificially breaking each read into fragments with a single SNP.
Example 2: GWAS for Major Depression (MD) in Humans
Exaample 2: GWAS for Major Depression (MD) in Humans

- Most human populations have much greater haplotypic diversity than the CFW mice, so it is necessary to increase the number of founders \( K \).
- As part of the CONVERGE study of Major Depressive Disorder, we sequenced the genomes of 5,224 Chinese women with MD and 5,218 controls.
- Mean coverage of nuclear genome was 1.7x.
GWAS for Major Depression

rs35936514 (LHPP) have frequencies of 3% and 8% respectively, compared to 45% and 26% in the CONVERGE cohort. We considered whether successful mapping of MDD in the CONVERGE samples was attributable to the recruitment of a severe, more genetically determined form of the disease. We tested that hypothesis by looking within the CONVERGE cohort at a particularly severe, and more heritable form of MDD: melancholia. Prior research has suggested that MDD patients with melancholia have more impairing, recurrent episodes and that risk for MDD is higher in the co-twins of probands with the melancholic subtype than in those with non-melancholic MDD. This increase is greater in monozygotic than dizygotic twin pairs, as would be expected if the subtype were associated with greater genetic risk.

In the CONVERGE cohort, 85% of cases met the DSM-IV criteria for melancholia. We searched for a genetic association in 9,846 samples (4,509 cases and 5,337 controls) and identified the same two loci that exceeded genome-wide significance on chromosome 10. The genomic control inflation factor for melancholia was 1.069, and for 1000 was 1.014. Even though the sample for melancholia was smaller than for MDD, at the SIRT1 locus the significance of association was two orders of magnitude greater than for MDD (top SNP rs80309727, chromosome 10:69617347, MAF 45.2%, P \(2.95 \times 10^{-12}\)). Extended Data Fig. 3 shows the Manhattan plot, quantile–quantile plot and detailed views of the SIRT1 locus associated with melancholia. All SNPs with P values of association, with melancholia are listed in Supplementary Table 5. To determine whether the increased association might have arisen by chance, we generated an empirical distribution of odds ratios by randomly selecting 4,509 cases from the total set and re-analysing the association with each of the genome-wide significant variants. We found that the observed value lay on the 98.8th percentile at the SIRT1 locus, but at the 61.6th percentile at the LHPP locus (Extended Data Fig. 4).

Our results indicate that, as others have suggested, obtaining low-sequence coverage of a large number of individuals can be an effective strategy.
We imputed 10,442 human genomes:

First using a conventional reference panel approach, (Beagle) to perform GWAS (Na et al 2015 Nature).

Subsequently with STITCH, without a reference panel, setting $K = 40$, for confirmation.

In order to speed up computation, we exploited the fact that all SNPs on a given read are from the same haplotype, to reduce computational complexity from $O(K^2)$ to $O(K)$ ("pseudo-haploid approximation").

Accuracy was measured relative to 70 individuals genotyped by arrays and 10 samples sequenced at higher coverage.
Performance of STITCH in CONVERGE compared to the Illumina HumanOmniZhongHua-8 array (left) and 10X sequencing (right). Results are shown for STITCH (K=40, 38 pseudo-haploid iterations, 2 diploid iterations), Beagle (upper default, lower 3 iterations with reference panel), and findhap(maxlen=50000, minlen=500, steps=3, iters=4) for the first 10 Mbp of chromosome 20 for n=11,670 Han Chinese samples. Post-QC is SNPs with info>0.4 and HWE p-value > $1 \times 10^{-6}$. 
Imputation Accuracy in Humans as a function of $K$

<table>
<thead>
<tr>
<th>$K$ haplotypes</th>
<th>$R^2$</th>
<th>Time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.922</td>
<td>24.5</td>
</tr>
<tr>
<td>20 pseudo-haploid</td>
<td>0.920</td>
<td>10.6</td>
</tr>
<tr>
<td>20 read-unaware</td>
<td>0.910</td>
<td>9.4</td>
</tr>
<tr>
<td>30 pseudo-haploid</td>
<td>0.917</td>
<td>12.4</td>
</tr>
<tr>
<td>30</td>
<td>0.927</td>
<td>52.2</td>
</tr>
<tr>
<td>40 pseudo-haploid</td>
<td>0.920</td>
<td>16.5</td>
</tr>
<tr>
<td>60 pseudo-haploid</td>
<td>0.923</td>
<td>27.3</td>
</tr>
<tr>
<td>80 pseudo-haploid</td>
<td>0.925</td>
<td>42.2</td>
</tr>
<tr>
<td>100 pseudo-haploid</td>
<td>0.927</td>
<td>61.1</td>
</tr>
</tbody>
</table>

Performance of STITCH on the first 10Mb of human chr 20. pseudo-haploid approximation was applied to 38 iterations, followed by 2 diploid iterations. Times are for a 4-core machine, scaled per 0.5Mb.
Effects of Varying Coverage and Sample Size

CFW mice (K=4) and CONVERGE humans (K=40) using STITCH without a reference panel. Validation is using array data, with each value representing the average for common SNPs (allele frequency 5-95%). Downsampling samples and reads was performed at random.
Arabidopsis

Accessions of Arabidopsis thaliana
MAGIC Arabidopsis populations

- 476 MAGIC lines sequenced at 0.3x
- STITCH predicts genotypes at 98% concordance with SNP genotypes
Genomic Rearrangements in *Arabidopsis* Considered as Quantitative Traits


*Wellcome Trust Centre for Human Genetics, †Department of Statistics, and ††Department of Plant Sciences, University of Oxford, OX1 3RB, United Kingdom, ‡Department of Computer Science, Swiss Federal Institute of Technology in Zurich, 8092, Switzerland, §Department of Biology and ††††Center for Cell and Genome Science, University of Utah, Salt Lake City, Utah 84112-0840, **Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany, ‡‡Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens Kew, Ardingly RH17 6TN, United Kingdom, §§§John Innes Centre, Norwich NR4 7UH, United Kingdom, ***UMR INRA-Agrocampus Ouest-Université de Rennes 1, 35653 Le Rheu Cedex, France, †††The Sainsbury Laboratory, Norwich Research Park, NR4 7UH, United Kingdom, ‡‡‡European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom, §§§Department of Biology and Biochemistry, University of Bath, BA2 7AY, United Kingdom, ***Department of Population Genetics, Gregor Mendel Institute of Molecular Plant Biology, 1030 Vienna, Austria, and ††††UCL Genetics Institute, University College London, WC1 6BT, United Kingdom
Structural Variant Traits

- Divide genome into 10kb segments
  - 11,915 bins in Arabidopsis (120Mb)
- Count number of read mapping anomalies in each segment and each individual
  - Structural Variant Traits
  - 6 different measures of mapping anomalies
- Variation in an SV-trait indicates a Structural Variant

Figure 2: Genome-wide distribution of the variance for the trait "improperly paired reads," computed in 10-kb windows. The x-axis shows genomic position and the y-axis the variance of each trait vector scaled by its mean. Each vertical line corresponds to a window; those with SV-QTL are blue (cis) and red (trans). The centromeres are marked by brown horizontal bars.

Variation in Arabidopsis chromosome 4 as reciprocal transpositions linking 1.61 and 2.65 Mb (Fransz et al. 2000), and a 93 kb inverted transposition identified previously in a cross between Ler-0 and Col-0 (Wijnker et al. 2013), and found it was present in 12 MAGIC founders.

To validate further SVs, we compared our SV calls for the founder accession Ler-0 against two Ler-0 contigs (chr3:16.65–17.02 Mb, chr5:25.06–25.23 Mb) that were independently resequenced and manually reassembled (Lai et al. 2011), thereby constituting a gold standard for comparison. The chromosome 3 contig (Figure 4) is enriched in SVs (83 indels, 31.100 bp), consistent with our analysis: 42 SV-QTL sources (36 cis and six trans) are in this region, and four trans SV-QTL map into it. As would be expected, the sources of these SV-QTL are within gaps in the contig. Furthermore, alignment revealed two long-range SVs within the contig (a transposition and a duplication that align to chromosomes 4 and 2, respectively), which coincide with the
Transpositions in a Population

**Population Ancestor**

Original state (Reference)  
![Image of original state](source L) → ![Image of original state](sink M)  

After Transposition  
![Image of transposition](breakpoint)  

**Present-Day Population Sample**

**Reality**  

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="L" alt="Image of A" /></td>
<td><img src="M" alt="Image of A" /></td>
</tr>
<tr>
<td><img src="L(%CE%B1)" alt="Image of B" /></td>
<td><img src="M(%CE%B2)" alt="Image of B" /></td>
</tr>
<tr>
<td><img src="L(%CE%B1)" alt="Image of C" /></td>
<td><img src="M" alt="Image of C" /></td>
</tr>
<tr>
<td><img src="L" alt="Image of D" /></td>
<td><img src="M(%CE%B2)" alt="Image of D" /></td>
</tr>
</tbody>
</table>

**Appearance after alignment to reference**  

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="L" alt="Image of A" /></td>
<td><img src="M" alt="Image of A" /></td>
</tr>
<tr>
<td><img src="L" alt="Image of B" /> ![Image of B](large insert size) <img src="M" alt="Image of B" /> ![Image of B](improperly paired, unmapped reads)</td>
<td></td>
</tr>
<tr>
<td><img src="L" alt="Image of C" /> ![Image of C](large insert size) <img src="M" alt="Image of C" /></td>
<td></td>
</tr>
<tr>
<td><img src="L" alt="Image of D" /> <img src="M" alt="Image of D" /> ![Image of D](improperly paired, unmapped reads)</td>
<td></td>
</tr>
</tbody>
</table>
Structural Variant QTLs

- Population correlations between SV-trait at a locus and genotypes at another locus
- 6502 SV-QTLs
- 25% transpositions (trans-SVQTLs)
Trans Structural Variation

source and sink of two trans SV-QTL mapped within the contig. Similarly, in the chromosome 5 contig, six cis SV-QTL correspond to deletions (Figure S3).

We also analyzed an independent de novo assembly of Ler-0 built from long PAC-BIO reads, GenBank accession GCA_000835945.1 (Berlin et al. 2015) to validate our trans SV predictions. This assembly was constructed algorithmically without manual revisions, and so is not guaranteed to be correct. Further, the Ler-0 individual sequenced in the PAC-BIO assembly was different from the individual that founded the MAGIC population, and therefore might carry private structural variations. Nonetheless, we expect it to be more accurate and contiguous than a Ler-0 assembly built from short Illumina reads alone. We took those 3080 Illumina paired-end reads for Ler-0 from Gan et al. (2011) that carried large insert size mapping anomalies when mapped to TAIR10, and that mapped to the sources of our predicted Ler-0 trans SV-QTL, and then mapped them to the PAC-BIO assembly using BWA (Li and Durbin 2010). These Illumina reads are from an individual grown from the same batch of seeds used to found the MAGIC population in /C24 2007, and should therefore share the same structural variants. Read anomalies with correct SV predictions should map contiguously to the PAC-BIO assembly, if the latter assembly accurately portrays the Ler-0 genome.

We found 2460 (80%) of these formerly split Illumina read pairs now mapped contiguously, defined as both members of a read-pair mapping to the PAC-BIO assembly with an insert size below 600 bp.

With the exception of these manually assembled Ler-0 contigs, and the provisional Ler-0 PAC-BIO assembly, the MAGIC founders are not contiguously reassembled into a genome-wide gold standard reference panel. Nevertheless, they provide information to test our predictions: at each SV-QTL, we predicted which founder haplotypes carried SVs at the origination of the population, under the assumption the SV was biallelic. Using the low coverage data for the 488 MAGIC lines, at each SV-QTL, we then predicted which founders carried the SV allele, based on correspondence between their SV-trait value and predicted founder allele, using the fact that SV haplotypes have elevated anomalous reads. We did this confidently at 2391 SVs where the founders partitioned into two groups, the remainder having complex multi-allelic SV predictions (Materials and Methods). Examples of founder partitions for cis and trans SV-QTL are shown in Figure S1.

We then examined the independently collected high-coverage reads in each of the 19 MAGIC founders (Gan et al. 2011) for read-mapping signatures that supported the predicted grouping of founders at each SV. We counted the read pairs linking source and sink at each of the 2391 SVs in the 19 high coverage founders. At 1585/2391 (66.3%, FDR 7.5%) SVs, we observed significant differences in anomalies between the predicted groupings of founders (Figure S4, which also shows that the majority of SVs were mapped within 50 kb). In the founders, the mean SV allele frequency was 6/19 = 31%. Only 387 (12%) were private to a single founder (Figure S2), in contrast to the fraction of SNPs (45%).
Structural Variation and Germination

Variation in the number of unpaired reads at 9.50–9.51 Mb explains 18.3% of the variance in resistance, and is adjacent to a cluster of leucine-rich repeat genes, and the genes RPP4 (Van Der Biezen et al. 2002), BAL (Yi and Richards 2009), and RPP5.

This locus is rearranged in some Arabidopsis accessions, and is known to affect disease resistance (Yi and Richards 2009); Figure 6 confirms the founder genomes have complex, polymorphic, SVs in this region. Since the resistance QTL is not completely ablated by the SV traits associated with it, additional nonstructural variants likely contribute to it.

Importantly, Figure 5B and Figure 6B show that correlations between SV traits and phenotypes are tightly localized, generally within the width of a single SV trait window, in contrast with wider linkage disequilibrium decay seen in QTL genetic mapping (Figure 5A). Therefore, correlations between SV traits and physiological traits pinpoint causal variants within physiological QTL that are otherwise too broad to localize [mapping resolution in MAGIC is ~200 kb (Kover et al. 2009)].

We also corroborated studies (Yalcin et al. 2011; Quadrana et al. 2016) showing SVs associate with gene dysregulation,
Structural Variation and Fungal Resistance

Even when the gene sequence is undisturbed. Within those SVs with mapped breakpoints, 119 genes spanned the breakpoints, 6909 lay inside the SVs (Table S8), and 21,747 outside. Using RNA-seq from 200 MAGIC aerial seedlings, scaled expression variance increased among genes spanning breakpoints (t-test: \(P < 0.001\)) and within SVs (\(P < 0.001\)) (Figure 7A). Similarly, more lines exhibited silenced transcripts for genes spanning breakpoints (t-test: \(P < 0.001\)), or within SVs (\(P < 0.001\)) (Figure 7B).

Expression within SVs was more correlated with local SV traits than outside SVs (F-test: \(P < 0.001\)) (Figure 7C).

Finally, we treated the SV traits as if they were quantitative, noisy genotypes, to compute pairwise correlations between MAGIC lines, as weighted correlations of their SV traits (Materials and Methods). We constructed SV genetic relationship matrices \(K_{SV}\); which we used to compute the SV-heritability \(h^2_{SV}\) of each of the physiological traits mapped above by analogy with the mixed models used for estimating SNP-based heritability (Kang et al. 2008a). This idea resembles the use of gene expression data to model intersample relationships (Kang et al. 2008b). We also compared these...
Structural Variation and Gene Expression

SV-heritabilities with those obtained from "classical" haplotype KH or SNP-based KSNP GRMs (Table 2). KH was computed from the identity between haplotype mosaics (i.e., IBD), while KSNP and KS were computed from the correlations of 1.2 M imputed SNPs or 12k SV-traits, respectively (Materials and Methods). We also computed SV heritability when only the most variable 50 or 25% of SV-traits were included, to test if heritability was concentrated at the most structurally variable loci.

As expected, SNP-based heritability $h^2_{SNP}$ resembles haplotype-based heritability $h^2_{H}$ for all phenotypes tested. However, the heritability $h^2_{SV}$ captured by the six measures of SV anomaly is more variable, sometimes being close to zero, but sometimes exceeding classical heritability considerably (Table 2). The SE of $h^2_{SV}$ was typically about twice that of $h^2_{SNP}$ or $h^2_{H}$; ($C^2_0$ compared to 0.05), presumably reflecting greater uncertainty in SV-traits. Therefore the larger $h^2_{SV}$ estimates should be treated with caution. Nonetheless, for phenotypes such as time to germination or bolting, the SEs of all estimates are $C^2_0$, and it is possible to compare them.

Figure 8, A and B illustrates likelihood curves for the times to germination (A) and bolting (B), for SNP, haplotype and large insert-size anomalies. Visualizing the entire curves gives a

![Figure A](image1.png)

![Figure B](image2.png)

![Figure C](image3.png)

Table 2 Estimates of heritability

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>$h^2_{H}$</th>
<th>$h^2_{SNP}$</th>
<th>$h^2_{SV}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resistance (resistance to A. laibachii)</td>
<td>0.000 (0.139)</td>
<td>0.258 (0.085)</td>
<td>0.490 (0.335)</td>
</tr>
<tr>
<td>RosetteLeafNumber.LongDay (number of leaves in a rosette for plants grown under long daylight)</td>
<td>0.228 (0.081)</td>
<td>0.322 (0.076)</td>
<td>0.463 (0.148)</td>
</tr>
<tr>
<td>RosetteLeafNumber.ShortDay (number of leaves in a rosette for plants grown under short daylight)</td>
<td>0.038 (0.060)</td>
<td>0.047 (0.062)</td>
<td>0.000 (NA)</td>
</tr>
<tr>
<td>Bolting.Bath (bolting time in a greenhouse)</td>
<td>0.426 (0.064)</td>
<td>0.476 (0.048)</td>
<td>0.783 (0.093)</td>
</tr>
<tr>
<td>Days.to.germ. (germination time)</td>
<td>0.220 (0.068)</td>
<td>0.149 (0.063)</td>
<td>0.385 (0.116)</td>
</tr>
<tr>
<td>FieldFT.pl (flowering time in the field)</td>
<td>0.000 (0.068)</td>
<td>0.095 (0.076)</td>
<td>0.000 (0.179)</td>
</tr>
<tr>
<td>FieldRD.pl (rosette diameter plasticity)</td>
<td>0.000 (NA)</td>
<td>0.000 (0.063)</td>
<td>0.000 (0.085)</td>
</tr>
<tr>
<td>Leaves.day.28.given.days.to.germ (residuals for number of leaves at day 28 regressed on germination)</td>
<td>0.193 (0.081)</td>
<td>0.299 (0.066)</td>
<td>0.391 (0.146)</td>
</tr>
<tr>
<td>ttl_branch.BATH (total number of branches of plants)</td>
<td>0.106 (0.048)</td>
<td>0.196 (0.054)</td>
<td>0.276 (0.104)</td>
</tr>
</tbody>
</table>

$h^2_{H}$ is haplotype-based heritability. $h^2_{SNP}$ is SNP-based heritability. $h^2_{SV}$ is the heritability estimated from structural variant anomaly traits. Numbers in brackets are the standard errors (SEs) of the heritability estimates above. Heritability for excess reads are not reported because the fraction of bins in any individual containing nonzero entries was too small. IP, Improperly-paired; LIS, Large Insert Size; SS, Same Strand; U, Unpaired; U + LIS, Unpaired or Large Insert Size.

1438 M. Imprialou et al.
Application to Other Species....

- Can this be applied to other species and populations?
  - rice MAGIC, accessions
  - chickpea MAGIC, accessions
  - pigionpea, sorghum, millet ....

- What about population structure?

- Do we need to sequence the MAGIC founders?
  - founders are like knowing the reference panel
  - but we can estimate them from the population
  - can identify introgressions?
Acknowledgements and References

- Funding provided by the Wellcome Trust, BBSRC
- CFW Mice:
  - Robbie Davies, Jerome Nicod, Jonathan Flint, Cai Na, MRC Harwell, Sanger Institute
  - Nicod et al 2016 Nature Genetics
  - Davies et al 2016 Nature Genetics
- CONVERGE Humans:
  - CONVERGE Consortium Cai Na, Jonathan Flint, Ken Kendler, BGI
  - Na et al 2015 Nature
  - Na et al 2015 Current Biology
- Arabidopsis Martha Imprialou, Paula Kover, and others
  - Imprialou et al 2017 Genetics