Summer Institute in Statistical Genetics 2014

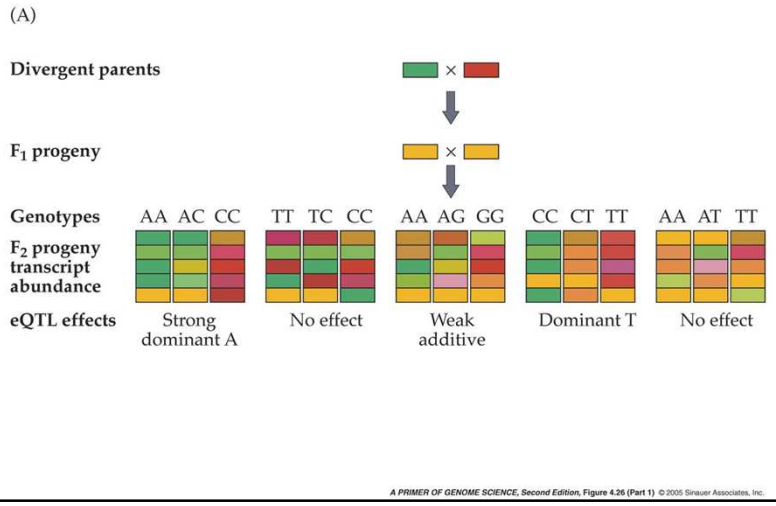
Gene Expression Profiling 5a
Genetics of Gene Expression: eSNPs

ggibson.gt@gmail.com
http://www.gibsongroup.biology.gatech.edu

Expression QTL analysis

- The architecture of transcription maps genotype onto phenotype
- Expression QTL (eQTL) are QTL that modulate transcript abundance in pedigrees or crosses
- It is estimated that at least 10% of transcripts differ in abundance between any two strains of most organisms; as much as 50% across a species
- Estimates of heritability of transcription also suggest that it is remarkably high, with transcription often showing a higher genetic component than visible traits

Principle of eQTL analysis

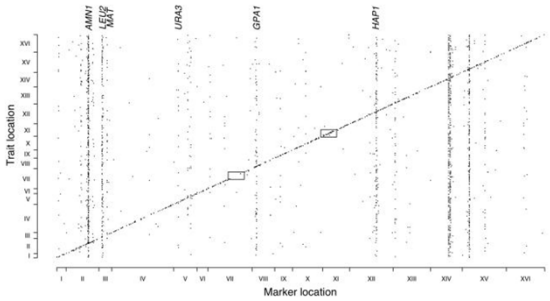


cis and *trans* eQTL

Schadt, Friend et al (2003) *Nature* **422**: 297-302

- Liver samples from 111 F2 mice from an obesity cross
- 15% of 23,500 genes with at least one eQTL explaining
  ~ 25% of the variance
- Tendency for strong eQTL to be in *cis* to the actual gene
- eQTL clustered in 7 hotspots (each 0.2% of the genome but >1% of the eQTLs)

Similarly for yeast:
Ronald and Akey,
***PLoS ONE* (2007) e678**

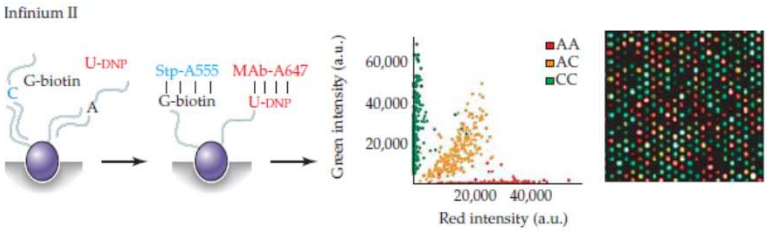


Limitations of eQTL analysis

- Any QTL experiment is only a comparison of two lines, so does not say anything about the frequency of QTL effects in a population
- If the number of F2 or BC progeny is less than 100, QTL analysis is prone to false positives, particularly for *trans*-hotspots
- Consequently, significance must be evaluated by permutation *being sure to permute the full genotype matrix against the full transcript abundance profile to preserve correlation structure*
- Resolution of QTL analysis is generally low (5 cM ~ 100-1,000genes), although enrichment for cis => most will be in the gene itself
- With pedigree analyses, ensure that one family is not driving the entire experiment

Principle of eSNP analysis

- Whole genome genotyping of >100 unrelated individuals



- Whole transcriptome profiling of the same individuals
- GWAS (Genome-wide association study) for transcription -> precise localization of regulatory SNPs in *cis* and *trans*

Significance thresholds

- Bonferroni for *cis*-linkages:
  0.05 / (20,000 genes x 250 SNPs) = 1 x 10^{-8}
- Permutation for *cis*-linkages:
  Random sets of n SNPs from distribtion of 2Mb windows
- Bonferroni for *trans*-linkages:
  0.05 / (20,000 genes x 500,000 SNPs) = 5 x 10^{-12}
- Permutation for *trans*-linkages:
  Randomize complete genotype and transcript matrices

OR adopt FDR criteria, although power not generally an issue
AND consider step-wise regression to adjust for LD

Gutenberg Heart Study example

Zeller et al (2011) *PLoS ONE* **5**: e10693

Significance level	Minimum R^2 [$]	Total number of associations	*cis*/*trans* ratio for associations	Total number of associated expressions (eQTLs)	*cis*/*trans* ratio for eQTLs	Total number of associated SNPs (eSNPs)	*cis*/*trans* ratio for eSNPs
$<10^{-6}$	0.016	93491	2.1	8575	0.5	67190	2.4
$<10^{-8}$	0.022	54749	7.3	3857	3.0	41425	11.2
$<10^{-10}$	0.028	42421	9.8	2998	6.0	33339	16.3
$<5.78\times10^{-12}$	**0.031**	**37403**	**10.7**	**2745**	**7.1**	**29912**	**17.1**
$<10^{-15}$	0.042	27330	12.7	2180	9.5	22591	17.8
$<10^{-20}$	0.057	19655	14.7	1725	12.8	16883	19.2
$<10^{-25}$	0.071	15015	16.4	1429	16.2	13045	21.5
$<10^{-35}$	0.099	9673	17.1	1031	21.6	8516	22.9
$<10^{-50}$	0.140	5873	14.0	712	28.8	5224	21.7
$<10^{-100}$	0.263	1790	10.5	290	28.1	1598	11.1
$<10^{-150}$	0.371	922	5.5	156	21.4	772	5.9
$<10^{-200}$	0.463	635	3.7	97	15.3	504	3.9
$<10^{-300}$	0.606	321	1.7	38	11.7	213	1.7

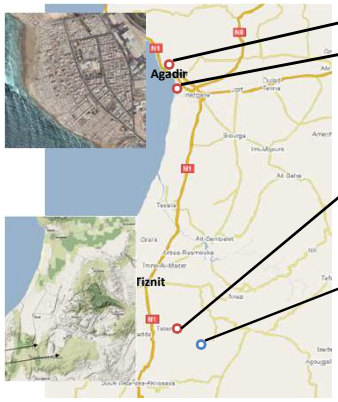
[$]*Minimum R^2 (proportion of gene expression variability explained by a SNP) observed for a given significance level. Numbers corresponding to study-wise significance are shown in bold. For investigating cis associations or performing any other hypothesis-based test, lower levels of significance may be considered.*
doi:10.1371/journal.pone.0010693.t002

Repeatability with GHS

	Stranger et al.		Dixon et al.		Schadt et al.	
Level of signifi-cance	Number of eQTLs at level of significance	Percent signifi-cant in GHS*	Number of eQTLs at level of significance	Percent signif-icant in GHS*	Number of eQTLs at level of significance	Percent signifi-cant in GHS*
$>10^{-8}$	86	55.8	110	50.9	928	47.9
10^{-8}–10^{-10}	63	69.8	162	50.0	168	57.7
10^{-10}–10^{-15}	144	63.2	237	54.8	211	57.3
10^{-15}–10^{-20}	60	70.0	102	60.7	120	66.7
10^{-20}–10^{-25}	38	89.5	73	65.7	73	67.1
$\leq 10^{-25}$	48	70.8	89	67.4	103	73.8
All	439	66.7	773	56.5	1603	54.1

** Comparisons were based on sets of gene expressions overlapping between each study and GHS and were restricted to autosomal cis eQTLs. All cis eQTLs considered significant in each study were retrieved and replication was assessed in GHS ($P<3.9\times10^{-6}$ correcting for 12,808 gene expressions).*
For Stranger et al [1], data were extracted from Table S2. We considered as significant the associations found in at least 3 HAPMAP populations. For Dixon et al [2], data were extracted from Table S1 and trans eQTLs were excluded. Matching of probes was done using a table provided by the authors on their web site. For Schadt et al [3], cis eQTLs considered significant (First.Pass.Indicator set to 1) were extracted from Table S3. For each eQTL, we selected in GHS the P-value of the best cis eSNP. The full data used to generate this table are provided in Files S2–S4.
doi:10.1371/journal.pone.0010693.t003

Moroccan experiment



Anza
Dcheira –
Mixed urban slums

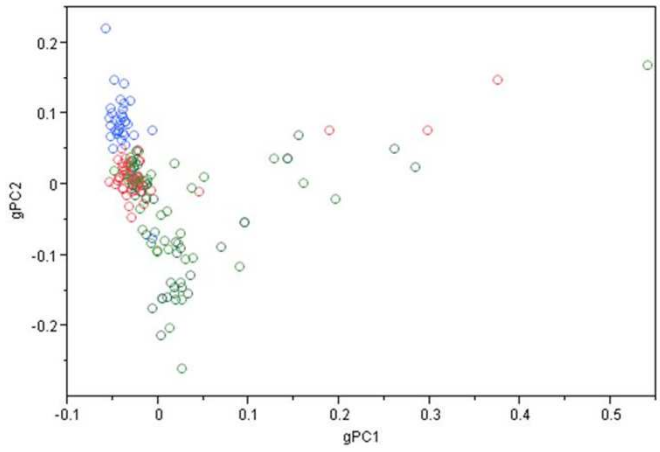
Ighrem –
Arab village

Boutroch –
Amazigh village

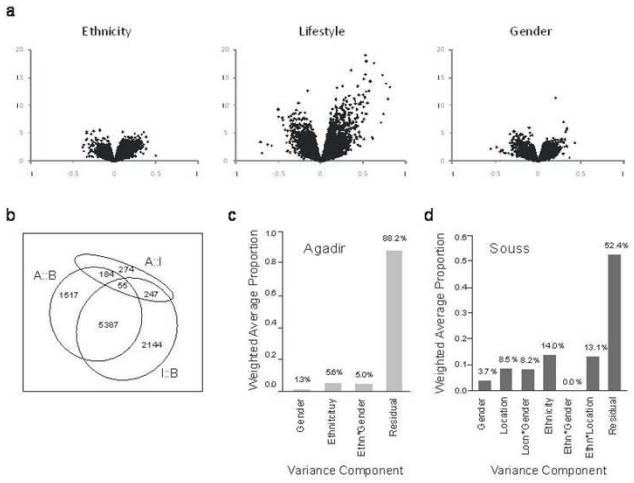
Gene Expression **mRNA**	Genomwide Genotyping **DNA**
HumanHT-12 **48,000 transcript probes**	Human 610-Quad **620,900 SNPs/CNV**

illumina

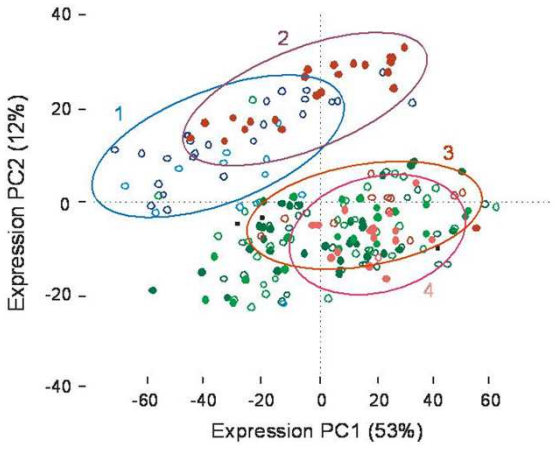
Visualizing population structure



Variance components



Cultural influences on expression



Association Dataset

genotypes_2219_recgeno

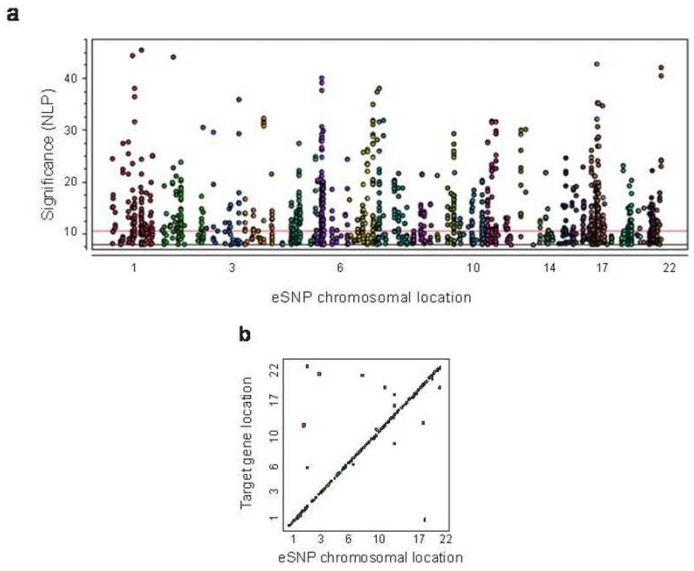
genotypes_2219_rec

Columns (1922/0)
Individual
Location
Lifestyle_2
Gender_2
Group
gPC1
gPC2
gPC3
recgeno1
recgeno2
recgeno3
recgeno4
recgeno5
recgeno6
recgeno7
recgeno8
recgeno9
recgeno10
recgeno11

Rows
All rows 194
Selected 0
Excluded 0

	Individual	Location	Lifestyle_2	Gender_2	Group	gPC1	gPC2	gPC3	recgeno1	recgeno2	recgeno3	recgeno4	recgeno
1	A03F	A	Urban	F	GP1	-0.0123	0.0079	0.0118	A/A	A/B	A/A	A/A	A/A
2	A09M	A	Urban	M	GP1	-0.0074	0.0142	0.0225	A/A	A/A	A/A	A/A	A/A
3	A103M	A	Urban	M	GP1	0.0904	-0.1153	0.0322	A/A	A/A	A/B	A/A	A/A
4	A105M	A	Urban	M	GP1	-0.027	0.0239	0.0266	A/A	A/A	A/A	A/A	A/A
5	A106M	A	Urban	M	GP1	-0.0281	0.0066	0.0365	A/A	A/B	A/A	A/A	A/A
6	A108M	A	Urban	M	GP1	-0.0308	0.039	0.0167	A/A	A/B	A/A	A/A	A/A
7	A109F	A	Urban	F	GP1	-0.0254	0.0318	0.0005	A/B	A/A	A/A	A/A	A/A
8	A110M	A	Urban	M	GP1	0.0511	0.0114	-0.0081	A/A	A/A	A/A	A/A	A/A
9	A112M	A	Urban	M	GP1	0.0246	-0.1389	0.0271	A/A	A/A	A/B	A/B	A/B
10	A117M	A	Urban	M	GP1	-0.0236	0.0041	0.0261	A/A	A/A	A/A	A/A	A/A
11	A119M	A	Urban	M	GP1	0.0306	-0.105	0.0129	A/A	A/A	A/A	A/B	A/B
12	A122M	A	Urban	M	GP1	0.0097	-0.0388	0.0121	A/A	A/A	A/B	A/A	A/A
13	A127M	A	Urban	M	GP1	0.196	-0.02	0.0211	A/A	A/A	A/A	A/A	A/A
14	A134M	A	Urban	M	GP1	-0.027	0.0239	0.0266	A/A	A/A	A/A	A/A	A/A
15	A135M	A	Urban	M	GP1	-0.0218	0.0472	0.0109	A/A	A/A	A/A	A/A	A/A
16	A136M	A	Urban	M	GP1	0.0074	-0.0791	0.0347	A/A	A/A	A/A	A/A	A/A
17	A138M	A	Urban	M	GP1	-0.0307	0.0253	0.0348	A/A	A/A	A/A	A/A	A/A
18	A139M	A	Urban	M	GP1	-0.0486	0.0189	0.0473	A/A	A/A	A/A	A/A	A/A
19	A142M	A	Urban	M	GP1	-0.0159	-0.0018	0.0251	A/A	A/A	A/A	A/A	A/A
20	A145M	A	Urban	M	GP1	-0.0168	-0.0642	-0.0097	A/A	A/A	A/A	A/A	A/A
21	A147M	A	Urban	M	GP1	0.022	-0.0835	0.0288	A/A	A/A	A/A	A/A	A/A
22	A150M	A	Urban	M	GP1	-0.0235	0.0068	0.0197	A/A	A/A	A/A	A/B	A/B
23	A18F	A	Urban	F	GP1	0.0264	-0.2605	0.0393	A/A	A/A	A/A	A/A	A/A

eSNP plots



Linear modeling

Simple association:

Expression = μ + SNP + ε

Adjusted for fixed covariates:

Expression = μ + Location + SNP + SNP* Location + ε

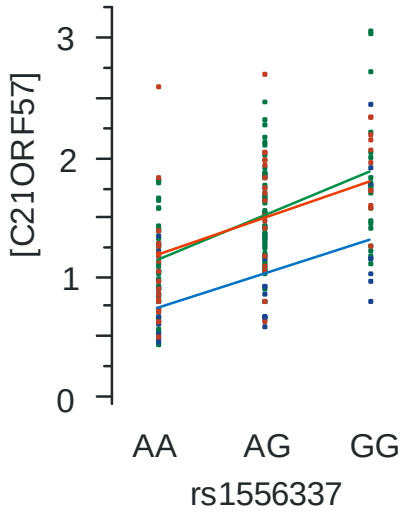
Adjusted for random and/or continuous covariates:

Expression = μ + Relatedness + Ethnicity + SNP + ε

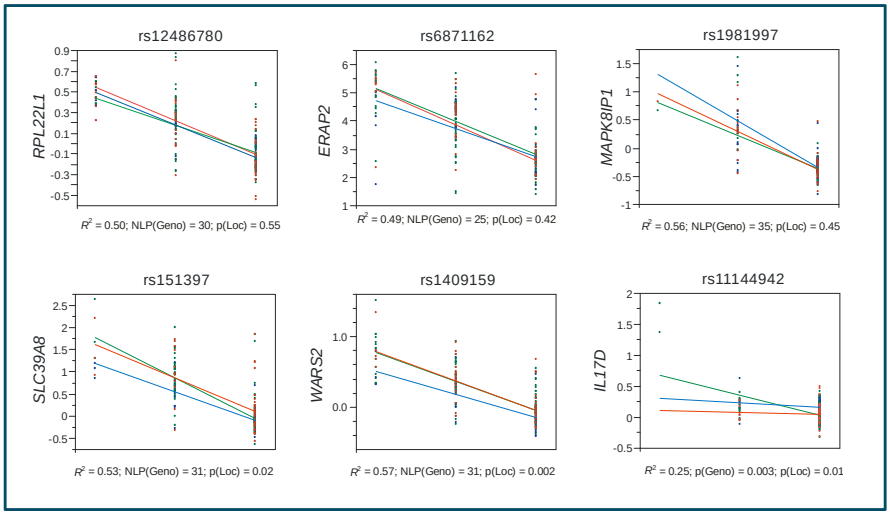
Alternate strategy to control for outliers if MAF < 5%:

Estimate Adjusted Expression Level, then perform SNP association on the rank order of the expression

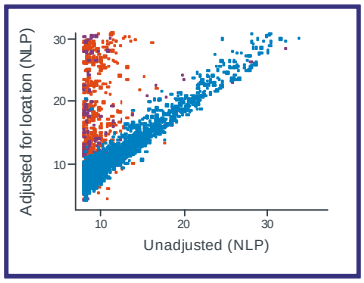
Additive Genotype & Environment

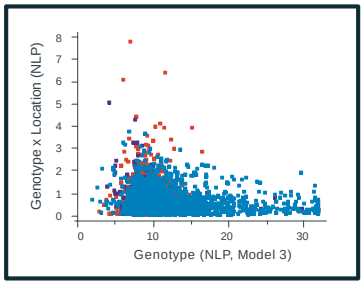


eSNP effect plots



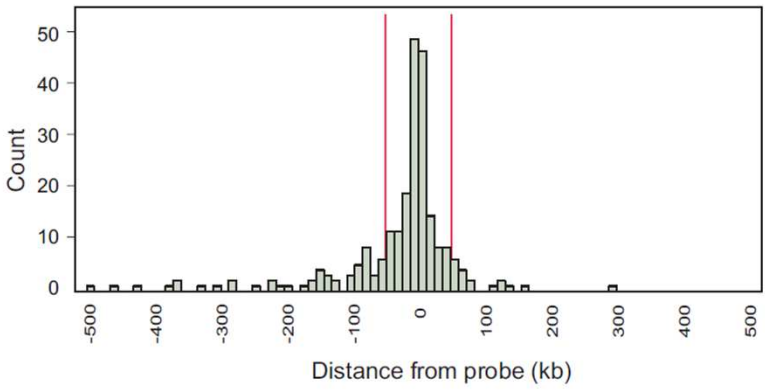
Absence of interaction effects





Red: Minor allele frequency < 0.05
Purple: 0.05 < MAF < 0.1

Location of eSNPs



Effect of Normalization

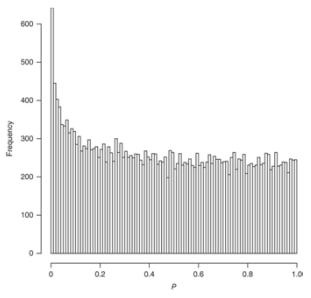
Table 3 **eSNP Analyses**

	Pearson Correlation			Spearman Rank Correlation		
Normalization	Total (NLP 8)	Cis (NLP 5)	Cis (NLP 8)	Probes (NLP 8)	Cis (NLP 8)	Probes (NLP 8)
RAW	552	1183	411	39	324	36
MEA	1082	2009	743	77	703	71
dr3	627	1362	455	44	407	46
DRM	959	2150	761	87	747	77
IQR	935	1708	603	71	565	73
LMN	484	1281	439	44	394	44
QNM	1211	2288	842	88	791	81
SNM	969	2084	825	86	821	81
PCA	602	1563	585	73	505	74

The Table reports the total number of associations detected between 34,548 Chromosome 6 SNPs and 732 Chromosome 6 Probes, respectively including total (trans and cis) associations at NLP 8; just cis associations at NLP 5 or NLP 8 (defining cis as eSNPs within 250 kb of the probe); the number of independent probes with eSNPs at NLP 8 (all using Pearson correlation with the transcript abundance); and then the cis associations and number of independent probes at NLP 8 using Spearman rank correlation.

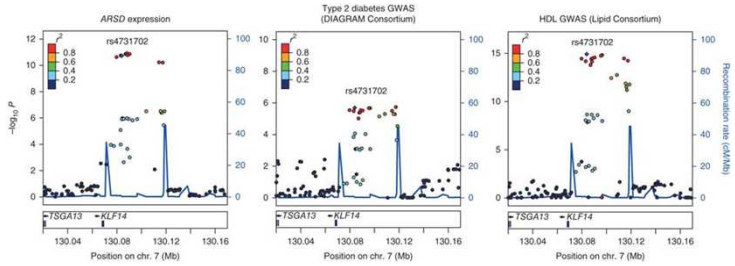
Trans-Effect of KLF14

Small et al (2011) *Nature Genetics* **43**: 561-564



		MuTHER		deCODE all		deCODE maternal		deCODE paternal		Combined MuTHER + deCODE maternal	
Gene	Chr.	Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Effect (s.e.)	P	Z score	P
AFAP1L1	15	0.08 (0.013)	1.2×10^{-9}	0.11 (0.059)	0.08	0.17 (0.085)	0.07	0.07 (0.083)	0.44	6.1	9.7×10^{-10}
ARSD	X	0.08 (0.012)	1.9×10^{-11}	0.24 (0.059)	2.2×10^{-4}	0.51 (0.083)	2.6×10^{-9}	−0.004 (0.083)	0.96	8.6	5.4×10^{-18}
C9orf62	9	0.09 (0.014)	4.8×10^{-10}	0.28 (0.058)	8.9×10^{-6}	0.59 (0.080)	2.1×10^{-14}	−0.09 (0.082)	0.28	9.3	1.1×10^{-20}
GNB1	1	0.05 (0.009)	4.0×10^{-8}	0.23 (0.059)	1.8×10^{-4}	0.42 (0.085)	1.6×10^{-6}	0.06 (0.084)	0.51	7.2	6.1×10^{-13}
KLF13	15	0.10 (0.017)	2.2×10^{-8}	−0.01 (0.060)	0.94	0.01 (0.086)	0.89	−0.02 (0.084)	0.80	4.8	1.4×10^{-6}
MYL5	4	0.09 (0.017)	4.5×10^{-8}	0.20 (0.059)	1.3×10^{-3}	0.45 (0.083)	1.3×10^{-7}	−0.04 (0.083)	0.60	7.4	1.1×10^{-13}
NINJ2	12	0.08 (0.013)	8.4×10^{-9}	0.14 (0.060)	0.03	0.24 (0.087)	0.01	0.05 (0.085)	0.59	6.3	4.1×10^{-10}
PRMT2	21	0.06 (0.010)	6.9×10^{-8}	0.18 (0.060)	0.01	0.27 (0.087)	6.7×10^{-3}	0.09 (0.085)	0.33	6.4	2.1×10^{-10}
SLC7A10	19	−0.27 (0.042)	2.7×10^{-10}	−0.21 (0.057)	7.4×10^{-4}	−0.31 (0.082)	3.3×10^{-4}	−0.11 (0.081)	0.18	−7.3	3.8×10^{-13}
TPMT	6	0.10 (0.013)	1.6×10^{-14}	−0.04 (0.060)	0.49	−0.03 (0.097)	0.78	−0.06 (0.084)	0.49	6.4	1.8×10^{-10}

The effect allele is the type 2 diabetes risk allele C, which has a frequency of 55% in the HapMap CEU population. Chr., chromosome; s.e., standar



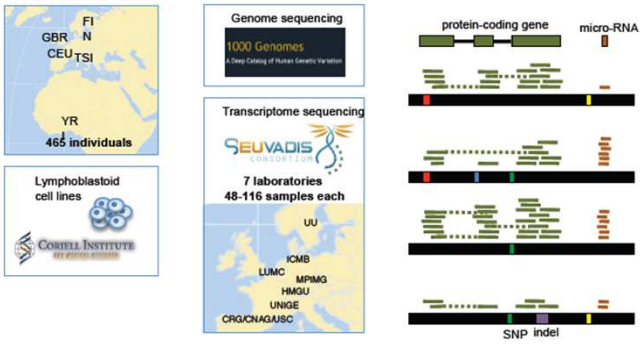
Challenges for eSNP analysis

- Great for finding transcripts regulated by one or two major effect SNPs that explain 20-60% of variance – but these are a minority
- Multiple comparison issues limit the power to detect weaker effects and to map several sites per transcript (unless N>10,000 ?)
- Outliers can produce very small p-values when MAF<5% and are quite common; PARTICULARLY with respect to interaction effects because one or two individuals will by chance be in a sub-group
- Only a few human tissues are accessible, and cost/ethics preclude recurrent sampling in many cases: hard to get longitudinal data
- Overlap between tissues estimated as only 10-20%, not much less than power to replicate 'marginal' associations at 10^{-8}

1000G eSNP study:

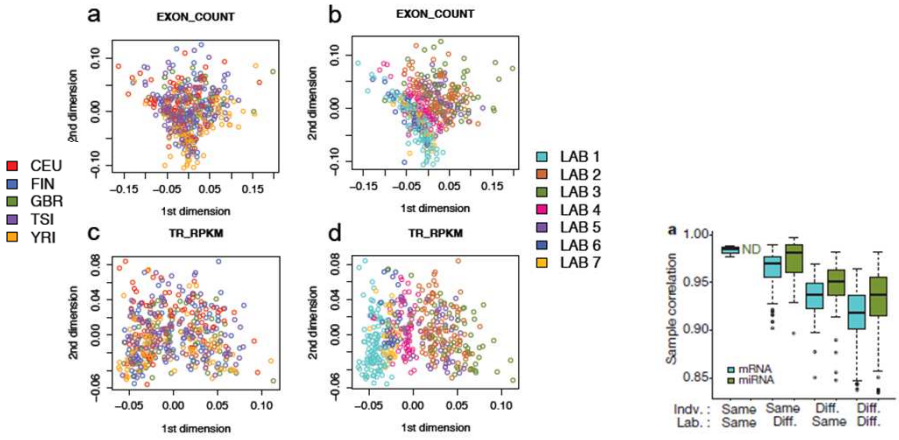
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

Performed RNA-Seq and miRNA-Seq on LCL for ~90 people each from five 1000G populations: Utah (CEPH), Finland, Britain, Tuscany and Nigeria (Yorubans)



Technical effects in the study

Sequencing in 7 laboratories showed inter-lab variance is less than among individual, yet there clearly are lab effects, particularly at transcript level

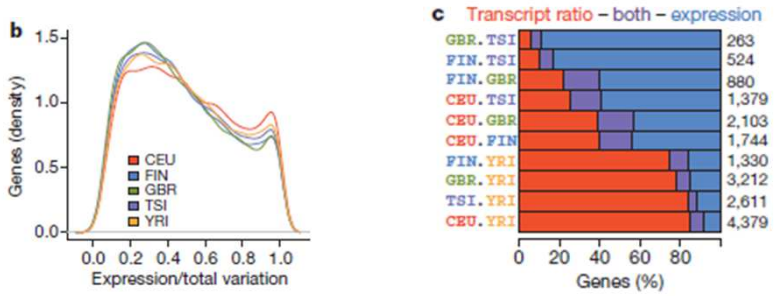


Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

Expression and Isoform components

Proportion of expression variation among individuals within populations ranges from 20% to 95%

Transcript ratio (isoform abundance) is greater than overall expression variation, and varies among populations, especially wrt Yorubans.



Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

eSNP analysis by RNA-Seq

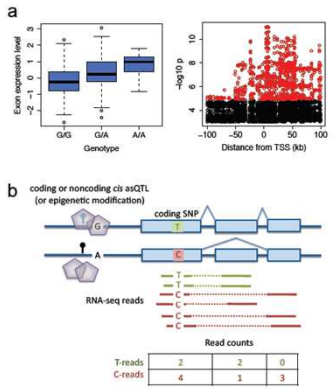
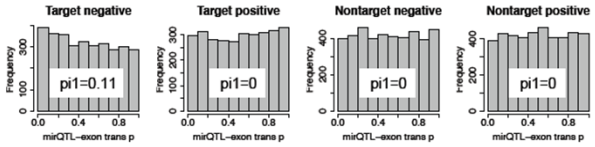


Table 1 | Numbers of transcriptome features with a QTL (FDR 5%)

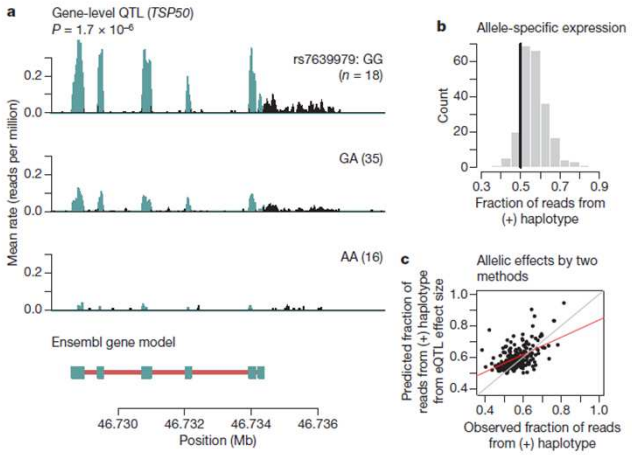
	Total	EUR ($n = 373$)	YRI ($n = 89$)	Union
Exon eQTL	12,981 genes	7,390	2,369	7,825
Gene eQTL	13,703 genes	3,259	501	3,773
Transcript ratio QTL	7,855 genes	620	83	639
mirQTL	644 miRNAs	57	15	60
Transcribed repeat eQTL	43,875 repeats	5,763	1,055	6,069

miR-QTL may also regulate negative targets



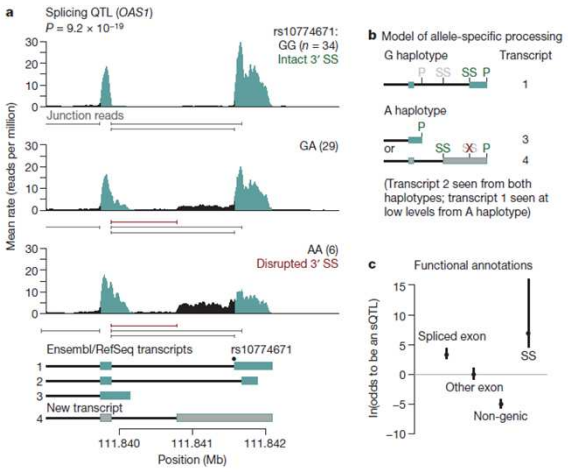
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

eQTL and Additivity



Pickrell et al. *Nature* **464**: 768-772 (2010)

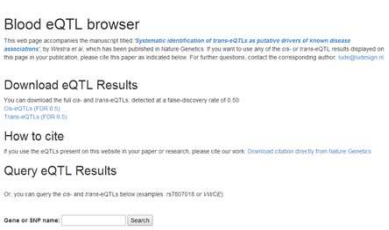
sQTL (Splicing QTL)



Pickrell et al. *Nature* **464**: 768-772 (2010)

Meta-analysis

http://genenetwork.nl/bloodeqtlbrowser/



eQTL meta-analysis on 5,311 individuals replicated in 2,775 more

Found trans-eQTL for 233 SNPs at 103 loci many of which are also disease QTL

Also generates local cis-eSNPs for almost half the genome

Summer Institute in Statistical Genetics 2014

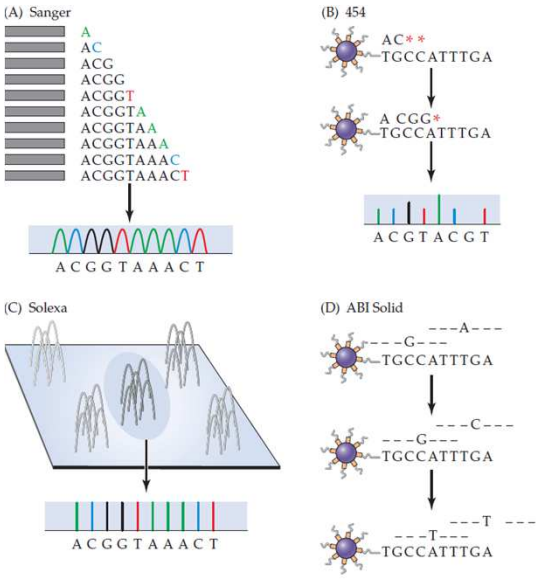
Gene Expression Profiling 5b
RNA-Seq

ggibson.gt@gmail.com
http://www.gibsongroup.biology.gatech.edu

Basic Workflow

1. Generate 40M 100+ bp reads, preferably paired-end or longer
2. Align to genome where available, or to close species,
   or assemble de novo transcriptome (Trinity, Velvet etc)
3. Realign to map ambiguous and/or intron-spanning reads
4. Perform QC, particularly looking for:
   duplicates, high number of mismatches, content biases
5. Infer abundance of genes and exons
6. Estimate isoform abundance by graph theory and paired reads
7. Perform differential gene expression analysis,
   adjusting for intensity-dependent variance

NextGen sequencing methods



The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Nagalakshmi, Snyder et al (2008) *Science* **320**: 1344-1349
Wilhelm, Bahler et al (2008) *Nature* **453**: 1239-1243

Single-nucleotide resolution transcriptome maps of both yeast genomes

75% - 90% of the nonrepetitive sequence of the yeast genome is transcribed.

High resolution detection of intron-exon splice junctions

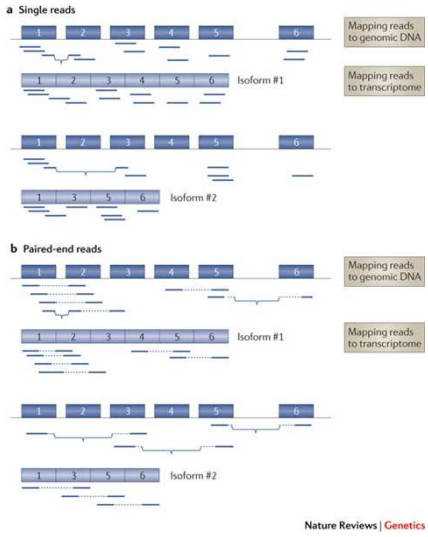
Correlation between splicing richness and transcript abundance

Unexpected complexity at 5' initiation sites including upstream ORFs

Unexpected heterogeneity at 3' ends, including opposite strand trancripts

Single and Paired-end reads

Ozsolak & Milos (2011) *Nature Reviews Genetics* **12**: 87-98



First mammalan RNASeq

Cloonan, Grimmond et al (2008) *Nature Methods* **5**: 613-619
Mortazavi, Wold et al (2008) *Nature Methods* **5**: 621-628

Embryonic stem cells and brain/liver/muscle

SQRLs are short quantitative RNA libraries, should be directionally tagged, and RNA fragmented prior to cDNA synthesis to avoid terminal biases

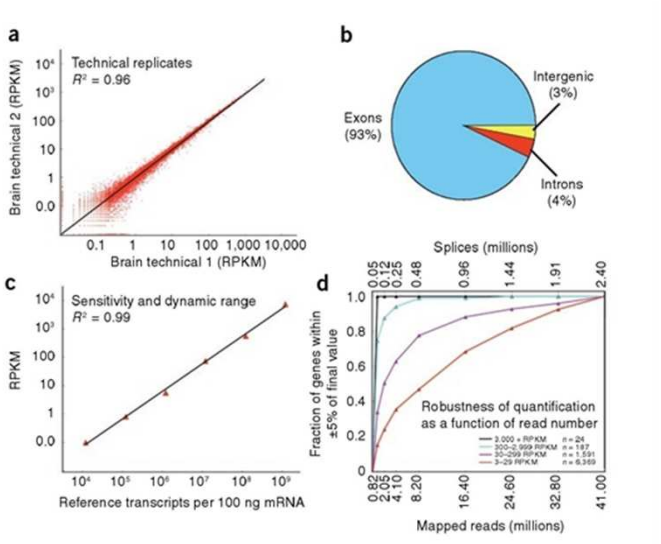
100 million high quality reads -> 11,500 expressed genes with >50 reads, estimate 40 million read required to reach saturation

Exons 50X enriched over introns, but 35% of reads outside known exons

Between 25% and 50% of SOLiD/Solexa reads are not uniquely mapable

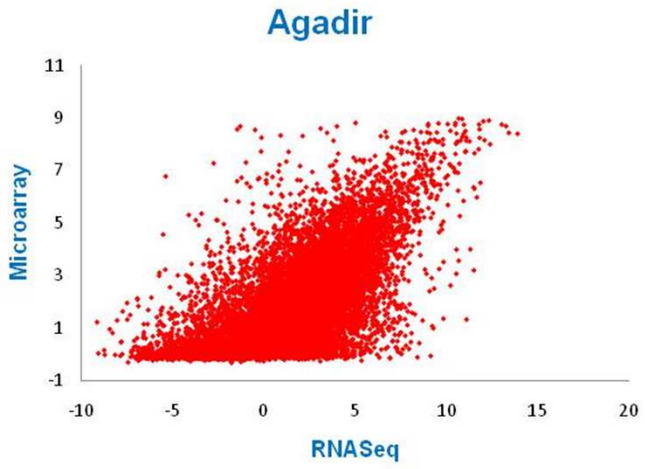
Note potential for measurement of allele-specific expression as well as high specificity for distinguishing among paralogs (not possible with microarrays)

Robustness, Linearity and Sensitivity



Mortazavi et al, 2008

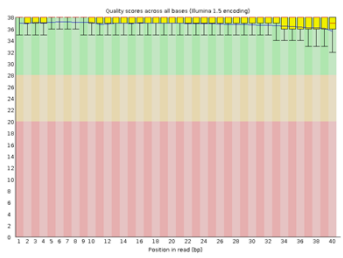
Accuracy Comparison



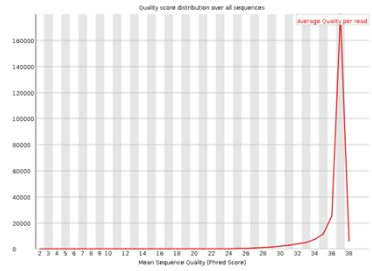
RNA-Seq Fast-QC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

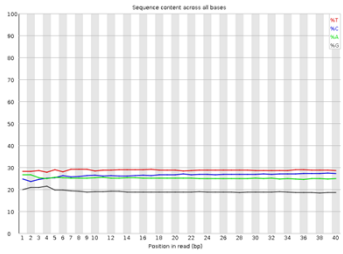
Per base Quality score



Per read Quality score



Per base nucleotide content



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2047	0.5178502762542756	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.4753496185060066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1646	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1641	0.4657363744918599	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1636	0.4644714739328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1631	0.4632065734330651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.4500516079415147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.4500516079415147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.4333549209690149 6	No Hit

http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

Contents [hide]

Fusion genes/chimeras/translocation finders/struc

Genome arrangements result of diseases like cancer can produce aberrant these modifications play important role in carcinogenesis studies.

- **BreakDancer** *BreakDancer*. See also *seqanswers/BreakDancer*.
- **ChimeraScan** *ChimeraScan*.
- **EBARDenovo** *EBARDenovo*.
- **FusionAnalyser** *FusionAnalyser*.
- **FusionCatcher** *FusionCatcher*.
- **FusionHunter** *FusionHunter* identifies fusion transcripts without depe aligner and paired-end reads. See also *seqanswers/FusionHunter*.
- **FusionMap** *FusionMap*.
- **FusionSeq** *FusionSeq*. See also seqanswers/FusionSeq.
- **SOAPFuse** *SOAPFuse*.
- **SOAPfusion** *Soapfusion*.
- **TopHat-Fusion** *TopHat-Fusion* is based on TopHat version and was require previous data about known genes and uses Bowtie to align cont
- **ViralFusionSeq** *ViralFusionSeq* is high-throughput sequencing (HTS transcripts at single-base resolution. See also *hkbic/VFS* and *SEQWil*
- **DeFuse** *DeFuse*.
- **PRADA** *prada*.

Tophat

Trapnell et al. *Bioinformatics* **25**: 1105-1111 (2009)

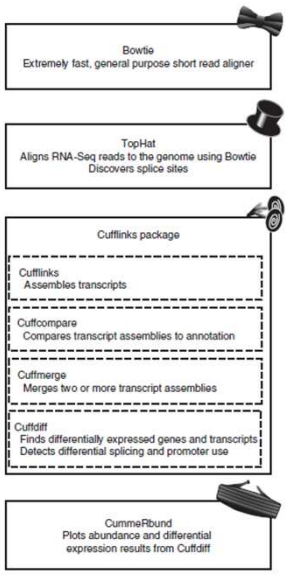
- Uses Bowtie to perform ultra-fast genomic read alignment
- Analyzes mappings to identify splice junctions
- Utilizes SAM specs for sequence alignment quality

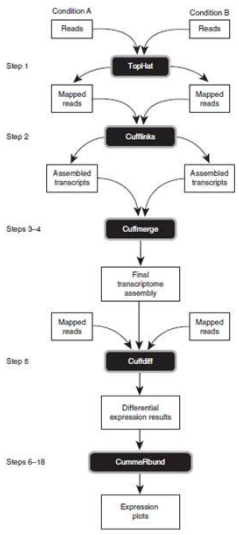
Cufflinks

Trapnell et al. *Nature Biotechnology* **28**: 511-514 (2010)
Trapnell et al. *Nature Protocols* **7**: 562-578 (2012)

- Assembles parsimonious transcript predictions
- Estimates their abundance
- Tests for differential expression

The Tuxedo Protocol





Trapnell et al, Nature Protocols **7**: 562-578 (2012)

Some other software tools

Bioconductor packages for RNA-Seq:

edgeR: empiricial Bayesian estimation for digital expression data

DEseq: differential expression analysis from RNA-Seq data

DEXSeq: differential expression of exons using GLM

rnaSeqMap: secondary RNA-Seq analysis with annotation

easyRNASeq: read count summarization and normalization

- Myrna: cloud-scale RNASeq analysis
- MISO, RSEM: maximum likelihood estimation algorithms
-Stampy, MapSplice, SpliceMap, GSNAP, Scripture: different aligners/parsers

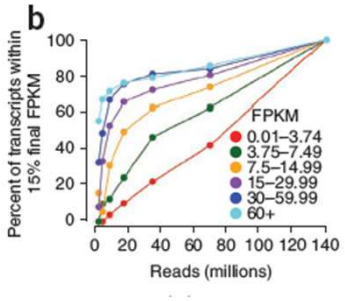
FPKM

Fragments per kilobase of transcript per million mapped fragments
(for paired-end reads; equivalent for single reads is RPKM)

- Adjustment "per kilobase of transcript" allows for absolute comparison of transcripts (or exons, or transcript isoforms)
- Adjustment "per million mapped fragments" allows for comparison of samples with variable quality and number of mapped reads

Transcript	Reads	Length	Mreads	FPKM
HoxA3	278	2.5 kb	27 million	4.12
	390	2.5 kb	38 million	4.11
Adh	956	1.7 kb	27 million	20.83
	831	1.7 kb	38 million	12.86
IL6	103	0.3 kb	27 million	12.72
	47	0.3 kb	38 million	4.12

Required read depth



At 50 million reads, 70% of transcripts with >7.5 FPKM are within 15% of estimated true abundance: ~ 5,000 transcripts

At 20 million reads, this requires FPKM > 15: only ~ 2,000 transcripts

Accuracy for FPKM < 5 is never particularly good: ~ 10,000 transcripts

Normalization issues

Statistical analysis of RNASeq data is where microarray analysis was 5 years ago: the broad outlines are in place, but there are few standards and the real analytical issues are just beginning to be appreciated.

In the FPKM formulation, the M stands for "million *mapped* reads" (not million reads) and hides the fact that read quality varies considerably from lane to lane, reflecting both sequence quality, and variation in library construction.

- Variance is density-dependent, which will affect hypothesis testing
- 5' to 3' biases arise due to variable reverse transcriptase efficiency
- Low quality libraries will lead to re-sequencing of individual amplicons
- GC content biases can affect alignment and amplification
- Repetitive DNA needs to be accounted for
- Low levels of intronic and intergenic reads complicate gene models/isoform prediction

Polymorphism will affect alignment, and lead to a tendency to overestimate the number of transcripts that map to one of the two alleles
Paired-end reads are very useful for reducing alignment mis-matches
Hard to estimate zygosity if there are fewer than 10 reads per individual

Alignment biases

Current standard is to deal with these by removing troublesome sites by:

- Genomic mapability score <1 (this is a UCSC track)
- Simulations of mapping polymorphic sites to HuRef19
- Failure to observe > 2% of alternate allele

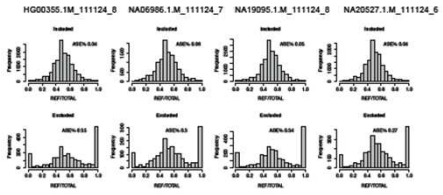


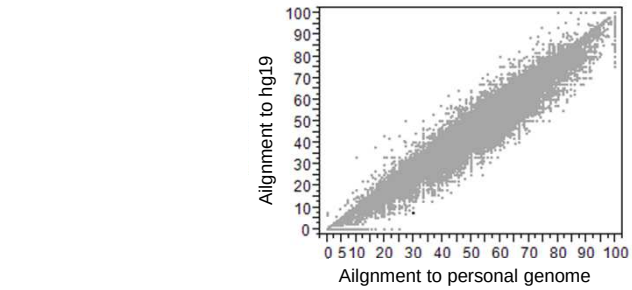
Figure S29. Filters for allelic mapping bias
Allelic ratio in four samples for SNPs that were kept in ASE analysis (top row) and in SNPs that were excluded due to increased risk of mapping bias based on simulations, genomic mapability estimates, or having only one allele observed (REF/TOTAL < 0.02 or > 0.98) (bottom row; see Supplementary Methods for details). The numbers denote the proportion of sites with significant (p<0.005) ASE, showing that excluded sites have clearly elevated ASE signal that is likely due to mapping problems.

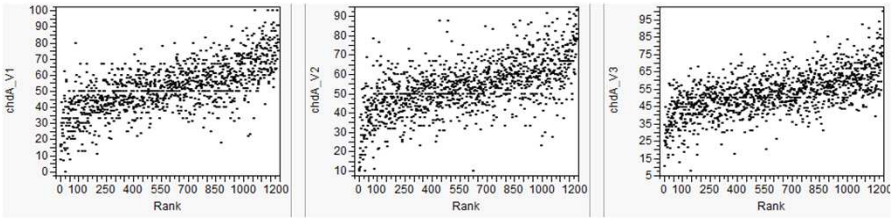
Lappalainen, Dermitzakis et al (2013) *Nature* **501**: 506-511

Variance of ASE in blood

Chromosome	Position	chdA_V1	chdA_V2	chdA_V3	chdB_V1	chdB_V2	chdB_V3	chdC_V1	chdC_V2	chdC_V3
chr14	60762929	63.6363636	88.8888889	54.5454546	60	66.6666667	57.1428571	41.1764706	25	35.2941177
chr15	34671438	53.3333333	60.7142857	68.9655172	•	•	•	62	63.7681159	60.7142857
chr15	39887649	•	•	•	21.4285714	42.8571429	16.6666667	53.1914894	53.9215686	51.3513514
chr15	41106485	•	•	•	71.4285714	53.8461539	66.6666667	51.8518519	52.3809524	55.5555556
chr15	67483979	•	•	•	65	60.8695652	61.5384615	33.3333333	46.6666667	40
chr15	79224747	71.1111111	60	61.8181818	52.9411765	44.6153846	54.1353384	50.8571429	46.2809917	41.1764706
chr15	79237247	46.1538462	41.1764706	55.8441558	•	•	•	47.3282443	56.1904762	62.992126
chr15	83781631	•	•	•	50	33.3333333	55.1724138	40	58.974359	39.6226415
chr16	8876202	•	•	•	68.75	62.5	75	56.25	37.5	44.6808511
chr16	8876880	•	•	•	78.2608696	76.4705882	85.1851852	48.6486487	40	65.625
chr16	24166130	42.5531915	47.4576271	49.0740741	•	•	•	55.6962025	54.8148148	56.445993
chr16	84883102	45.4545455	35	34.2105263	•	•	•	75.3846154	61.5384615	76.4705882
chr16	84922859	81.8181818	77.7777778	73.6842105	•	•	•	75.6906077	77.8947368	75
chr17	27959903	•	•	•	34.2465753	34.1463415	46.2616822	56.0209424	58.6956522	56.9444444
chr17	39084504	25.9259259	28.5714286	33.3333333	•	•	•	57.4712644	55.1111111	50.5434783
chr17	42449789	68.1818182	70	61.7647059	•	•	•	53.7572254	52.6427061	55.026455
chr17	72613589	45.4545455	56.6037736	59.5744681	•	•	•	39.2307692	40.2439024	33.3333333
chr17	78093353	68.0851064	67.0454546	72.8395062	45.9183674	46.3414634	46.4285714	56.4102564	58.8405797	64.2424242
chr19	1056065	52.173913	72.2222222	87.5	•	•	•	22.3684211	16.2393162	13.1578947
chr19	1064193	67.8571429	57.6923077	70.8333333	•	•	•	21.4285714	19.2546584	23.5294118
chr19	6919753	42.8571429	51.5151515	47.2972973	•	•	•	40.5405405	44.4444444	52.9411765
chr19	40947305	•	•	•	29.1666667	28.5714286	29.2682927	45.4545455	45.7142857	48.2758621
chr19	42083849	28.5714286	18.1818182	22.2222222	73.6842105	85.7142857	70	•	•	•
chr2	28865760	90.9090909	93.3333333	100	•	•	•	27.2727273	18.9189189	39.1304348
chr2	85570273	55.5555556	28.5714286	20.8333333	•	•	•	52.173913	50	44.2307692
chr2	85622059	•	•	•	45.7142857	50	54.9295775	58.5227273	52.014652	49.0322581
chr2	85625222	44.4444444	50.617284	60	•	•	•	46.0674157	51.369863	55.0877193

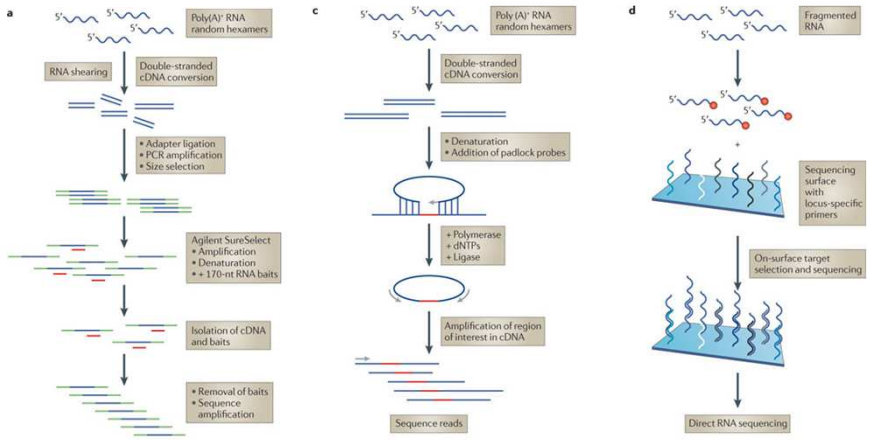
Self-genome alignment





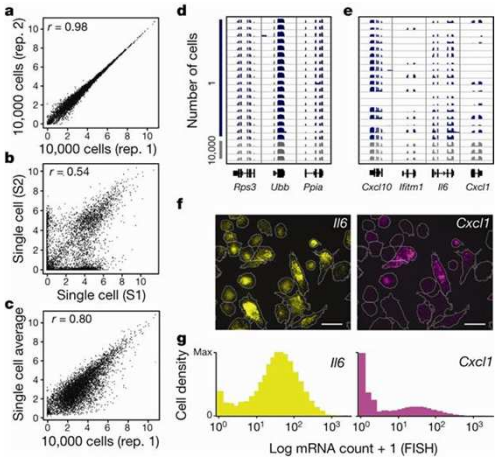
Targeted RNASeq

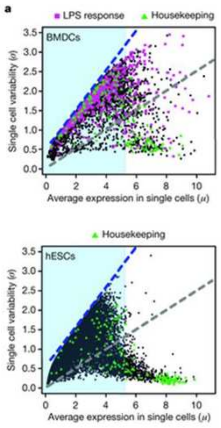
Ozsolak & Milos (2011) *Nature Reviews Genetics* **12**: 87-98



Single cell RNASeq

AK Shalek *et al.* (2014) *Nature* **498**: 236-240
Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells





nature